UNIVERSIDADE DE LISBOA

FACULDADE DE LETRAS

Qualidade na tradução automática e na pós-edição: anotação de erros de concordância e ordem de palavras

Rhandra Taysk da Silva Lopes

Dissertação orientada pela Prof.ª Doutora Anabela Proença Leitão Martins Gonçalves e pela Prof.ª Doutora Helena Gorete Silva Moniz, especialmente elaborada para a obtenção do grau de Mestre em Tradução

2019

AGRADECIMENTOS

Primeiramente, gostaria de agradecer a minha orientadora, Anabela Gonçalves, por ter me acompanhado neste caminho com toda a paciência e serenidade, dando auxílio nas questões mais difíceis. Este trabalho não teria sido possível sem as suas ideias e inúmeros conselhos valiosos.

Agradeço também à minha coorientadora, Helena Moniz, que me auxiliou enormemente durante a minha estadia na empresa de acolhimento e ofereceu excelentes comentários que ajudaram na construção da arquitetura da presente pesquisa.

Um grande obrigada às minhas famílias, na *Casa do Abacateiro* e na *Maison Jaune*, pelo apoio constante que recebi durante esse período. Não tenho palavras para expressar agradecer pela confiança que depositaram em mim e por me presentearem todos os dias com carinho e coragem.

Gostaria também de fazer um agradecimento especial ao meu companheiro de vida, tão presente com seus sorrisos durante todos esses momentos.

RESUMO

Considerando-se as características da tradução automática, como o baixo custo e a rapidez, esse tipo de tradução tem sido cada vez mais utilizado no mercado de tradução. Todavia, a qualidade dos resultados obtidos pelos sistemas utilizados pode não ser ideal, sendo necessário fazer a tradução passar por um processo de pós-edição, feita por humanos, para atingir níveis de qualidade satisfatórios. O presente trabalho procura descrever o processo de tradução automática, pós-edição e anotação oferecido pela plataforma Unbabel, que faz uso de uma *crowd* para a edição online dos erros encontrados nos textos traduzidos pelo sistema *Neural Machine Translation* (NMT). O objetivo principal da presente pesquisa é aprimorar a qualidade dos textos traduzidos por essa empresa, através de propostas de aperfeiçoamento das orientações fornecidas pela empresa aos seus editores e anotadores e através de sugestões para a avaliação e treinamento desses elementos humanos. Para atingir esse objetivo, foram coletados e analisados dados contendo trechos de textos traduzidos pelo sistema automático, pós-editados por humanos e anotados também por humanos sob as etiquetas de *Agreement* e *Word Order*, tendo o inglês como língua de partida e o português brasileiro como língua de chegada. A partir da observação dos resultados dessas análises, foi possível definir *Golden Texts* e testes de múltipla escolha com mensagens de *feedback* para auxiliar na avaliação e treinamento dos anotadores e pós-editores.

Palavras-chave: tradução automática, concordância, ordem de palavras, pós-edição, anotação de erros

ABSTRACT

Considering the characteristics of machine translation, as low costs and speed, it has been increasingly in the translation market. Nevertheless, the quality of the results obtained with these systems may not be ideal, a post-edition step, done by humans, to reach satisfactory quality levels being necessary. The present work describes the translation, post-edition and annotation done at Unbabel's platform, that uses a crowd to edit online the errors occurring in the texts translated by the *Neural Machine Translation* (NMT) system. The main objective of this research is to enhance the quality of the texts translated in that platform, through the suggestions of improvements in the guidelines the company gives to its editors and annotators and also through suggestions

to evaluate and train these people. In order to achieve this goals, data containing parts of texts translated by the machine translation, post-edited by humans and annotated, also by humans, under the *Agreement* and *Word Order* labels, was collected and analyzed. This data had the English as source language and Brazilian Portuguese as target text. The results of these analyzes make it possible to define *Golden Texts* and multiple choice tests to help the evaluation and training of annotators and post-editors.

Keywords: machine translation, agreement, word order, post-edition, error annotation

ÍNDICE

1. INTRODUÇÃO

O presente trabalho foi feito no contexto do mestrado em tradução da Faculdade de Letras da Universidade de Lisboa[1]. Para isso, foi feito um estágio na empresa de tradução automática Unbabel, que utiliza pós-editores humanos para melhorar a qualidade de suas traduções. O objetivo geral deste trabalho é fornecer instrumentos que possam auxiliar no aprimoramento da qualidade da tradução produzida pela Unbabel. O objetivo específico é melhorar a avaliação e o treinamento dos anotadores e pós-editores da Unbabel, principalmente os que trabalham com o par linguístico inglês-português brasileiro, objeto de análise na presente pesquisa.

A Unbabel combina a tradução automática e a pós-edição humana. Esses pós-editores, que não são necessariamente tradutores profissionais, editam textos traduzidos pelo sistema de tradução automática através de uma plataforma online. Por meio desse método, é possível aumentar o número de traduções produzidas e diminuir o custo, tendo em vista a rapidez e rentabilidade dos sistemas de tradução automática quando comparados aos tradutores humanos.

A área da Tradução Automática tem passado por um grande crescimento nas últimas décadas, conquistando cada vez mais importância no mercado de tradução e despertando o interesse de pesquisadores nas áreas computacional e linguística. Apesar dos avanços na elaboração de sistemas mais eficientes, a qualidade das traduções obtidas ainda é variável e não totalmente satisfatória. Por isso, esse tipo de tradução é frequentemente acompanhado por uma etapa de pós-edição do texto traduzido, através da qual os erros são corrigidos por editores humanos. Contudo, a pós-edição possui um custo elevado, além de ser um processo mais demorado do que simplesmente o sistema de tradução automática, sendo interessante para empresas como a Unbabel investir na melhoria desse processo.

A Unbabel já possui métodos de avaliação automática e humana da qualidade dos textos traduzidos e pós-editados. No entanto, esses métodos ainda necessitam de aprimoramentos para aumentar a uniformidade e a fiabilidade dos resultados. A presente pesquisa foca a avaliação feita por humanos que, tal como outras áreas, tem critérios de avaliação que poderão levantar dúvidas, tendo em vista as diferentes categorias de erro utilizadas na empresa serem complexas e poderem suscitar ambiguidades.

[1] A presente pesquisa foi escrita na variante brasileira do português.

1.1 Objetivos

O objetivo principal desta pesquisa é fornecer instrumentos mais esclarecedores à anotação, auxiliando assim no processo de avaliação da qualidade da tradução produzida na Unbabel. Os objetivos específicos são contribuir para melhores *guidelines*, mais detalhadas em função de categorias de erro complexas, bem como fornecer sugestões de avaliação e treinamento de anotadores e pós-editores envolvidos na edição e anotação de textos traduzidos do inglês para o português brasileiro (doravante PB). Para alcançar esses objetivos, será feita uma análise do processo de tradução, pós-edição e anotação feito na Unbabel, focando-se nas orientações dadas por essa empresa através de suas *Language Guidelines* e *Annotation Guidelines*.

O *corpus* de textos coletados[2] é constituído por trechos de textos traduzidos pela máquina, pós-editados por humanos e categorizados por outros humanos sob as etiquetas *Agreement* e *Word Order*, ambas bastante complexas dentro do sistema de catgorização utilizado na Unbabel, sendo o inglês a língua de partida e o PB a língua de chegada. A escolha dessas etiquetas se deve ao fato de os fenômenos a que elas correspondem estarem entre as categorias de erros mais frequentes dentro desse par linguístico e afetarem enormemente a compreensão e qualidade dos textos traduzidos.

Tendo em vista a intenção de analisar os erros de concordância e ordem de palavras ainda presentes nos textos escritos em PB após as etapas de tradução automática e pós-edição, buscou-se fazer uma descrição dos fenômenos de concordância e de algumas questões associadas a ordem de palavras, principalmente em casos mais problemáticos, nos quais há hesitações entre os próprios falantes nativos de PB. Também o processo de anotação foi observado pormenorizadamente através da análise das orientações gerais dadas pela empresa em suas *Annotation Guidelines*.

Esses processos nos permitiram identificar os tipos de erros mais comuns que persistem após a pós-edição. Através dessa identificação, fomos capazes de identificar os aspectos mais problemáticos que podem ser tidos em consideração pela empresa para melhorar a eficiência e a qualidade do processo de tradução. Também através dessa análise, foi possível formular propostas de avaliação e treinamento de pós-editores e anotadores. Essas propostas podem contribuir na assimilação das orientações dadas nas *Guidelines* da empresa e promover a maior uniformidade das anotações e a qualidade das pós-edições.

[2] Por razões de confidencialidade, o documento contendo a lista completa do *corpus* analisado na presente pesquisa será fornecido somente ao júri avaliador e não será partilhado ou guardado posteriormente.

1.2 Metodologia

A investigação do presente trabalho se iniciou com a revisão das principais teorias acerca do desenvolvimento da tradução automática, que permitiu o entendimento do sistema de tradução automática utilizado na empresa e a importância da etapa de pós-edição para a melhoria da qualidade da tradução feita pelo sistema. Durante essa etapa, foi possível verificar o papel da avaliação da qualidade da tradução e a dificuldade em aumentar a concordância inter-anotadores, tendo em vista a influência da subjetividade no processo de anotação de erros. Estudos como os de Comparin (2016) e Figueira (2018), feitos no contexto da Unbabel, demonstram que sugestões de melhorias nas *guidelines* e a implementação de *decision trees* são benéficas para esse processo. Esses dois trabalhos analisam diversas etiquetas utilizadas na anotação da empresa, mas não abordam profundamente as duas categorias que serão trabalhadas na presente pesquisa.

Durante o estágio, foi possível observar e utilizar as plataformas de anotação e pós-edição da empresa. Nesse momento, foram examinadas as duas principais orientações fornecidas pela empresa aos pós-editores e aos anotadores: a *Language Guidelines* e as *Annotation Guidelines*, respectivamente. Os dados coletados para a presente pesquisa são trechos de textos traduzidos através do sistema de tradução automática, do inglês para o PB, pós-editados por humanos e finalmente anotados também por humanos, todos, em princípio, falantes de PB como língua materna. Para a presente pesquisa, foram selecionados apenas os erros anotados sob as categorias de "*Word Order*" e *Agreement*". Após a recolha dos dados que serão objeto de análise na presente pesquisa, esses dados foram anonimizados, dada a necessidade de preservar informação confidencial, de acordo com o Regime Geral de Protecção de Dados, em vigor na União Europeia.

A descrição dos fenômenos de concordância e ordem de palavras em PB permitiu a análise dos erros e casos problemáticos encontrados nos dados. Essa descrição também foi utilizada no processo de elaboração de sugestões e pode ser utilizado como material pela própria empresa, principalmente em casos mais controversos, pois reúne pesquisas recentes acerca dos temas tratados.

As unidades selecionadas e categorizadas pelos anotadores da Unbabel passaram por uma anotação feita no contexto da presente pesquisa para permitir a comparação entre a anotação feita por esses anotadores e as orientações dadas nas *Guidelines* da empresa. A partir dessa comparação foi possível elaborar sugestões de treinamento e avaliação.

Também a partir dessa comparação, foi possível filtrar os dados que não continham erros ou que continham outros tipos de erros, não sendo objetos da presente pesquisa. Os dados que continham efetivamente erros ou casos problemáticos foram então analisados novamente sob uma perspectiva mais gramatical, procurando-se entender as estruturas mais envolvidas nos erros e a motivação por trás desses erros. A elaboração de sugestões para a avaliação e treinamento de anotadores e pós-editores foi possível graças aos resultados dessa análise.

1.3 Organização

Quanto à organização do presente trabalho, na seção 2 é apresentada uma perspectiva histórica da tradução automática, focando temas relacionados com o processo feito na Unbabel. Nessa seção, são apresentados os diferentes paradigmas de tradução automática, ressaltando-se o desenvolvimento do *Neural Machine Translation* (NMT), sistema utilizado pela empresa. Na seção 3, é descrita a plataforma Unbabel, descrevendo-se as etapas do processo de tradução e pós-edição. Na seção 4, é observado o processo de anotação da Ubabel, com ênfase no sistema de categorização de erros utilizado e nas *Annotation Guidelines*, objetos da presente pesquisa. Na seção 5, é fornecida uma descrição dos fenômenos de concordância e ordem de palavras (em particular, a ordem dos complementos clíticos e dos advérbios) na variante brasileira do português, com foco em aspectos mais problemáticos e fornecendo-se excertos de *corpus* escritos nessa variante para exemplificar os fenômenos descritos. Essa seção é crucial para o presente trabalho, tendo em vista fornecer informações importantes para a análise dos erros encontrados nos dados recolhidos na Unbabel e para a elaboração de sugestões. A seção 6 é dedicada à análise dos dados fornecidos pela empresa. Essa análise é feita em duas etapas, nas quais dois aspectos distintos são observados: primeiramente, são observados os erros de categorização e segmentação feitos pelos anotadores, ou seja, casos em que não são seguidas as instruções dadas nas *Guidelines* ou em que a anotação foi feita de forma não uniforme; na segunda parte, foram observados os erros ou casos problemáticos ligados estritamente aos fenômenos de concordância e ordem de palavras, procurando-se entender a origem desses problemas. Durante as duas etapas, foram assinalados os erros mais frequentes feitos por anotadores e pós-editores. Na seção 7, a partir da análise de erros feita na seção anterior, são apresentadas sugestões que podem ajudar a resolver casos problemáticos nas etapas de pós-edição e anotação, tendo em conta o par linguístico e os fenômenos envolvidos nos dados analisados. Também são

sugeridos processos de avaliação e treinamento de editores e anotadores, buscando assim melhorar a qualidade dos textos e anotações de erros feitos na empresa. Na última seção, são apresentadas as conclusões da presente pesquisa, sendo ainda explicitados temas e questões não respondidas na presente pesquisa, mas que podem ser objeto de trabalhos futuros.

2. TRADUÇÃO AUTOMÁTICA

Neste seção serão apresentados os principais elementos que auxiliaram na construção do enquadramento teórico acerca da tradução automática durante a presente pesquisa. Na seção 2.1, será fornecida uma visão geral dos momentos históricos que marcaram o seu desenvolvimento. Na seção 2.2, é descrito o sistema baseado em neurônios (*neural-based machine translation)*, inserido nos paradigmas de tradução automática baseados em *corpora* (*corpus based machine translation)*[3].

A expressão "Tradução Automática" se refere aos sistemas computadorizados responsáveis pela produção de traduções a partir de uma língua natural para outra, com ou sem o auxílio humano. Embora o anseio dos desenvolvedores de tradução automática ser a produção de traduções de alta qualidade, em alguns casos os textos traduzidos pelas máquinas são revisados por humanos, através da pós-edição (Hutchins e Somers, 1992: 3). Dorr et al. (1999: 2) também ressaltam que em geral a tradução automática não visa uma tradução completamente automática e de alta qualidade, mas sim um nível de tradução que se adeque às necessidades básicas do utilizador, talvez necessitando de um *input* controlado (pré-edição), da revisão do *output* (pós-edição) ou de ambos esses processos para chegar ao resultado final. Como ressalta Quah (2006: 43-44), esses dois processos não são sempre necessários, mas podem ser solicitados devido a inúmeros fatores, como a qualidade linguística do texto na língua de partida, o tipo de ferramenta de tradução utilizada e a qualidade exigida do texto de chegada.

2.1 Perpectiva histórica geral da Tradução Automática

Hutchins (2010: 1) afirma ser possível traçar as origens da tradução automática no século XVII com as ideias de linguagem universal e de dicionários mecânicos, mas é somente no século XX que as primeiras propostas práticas ocorreram. O desenvolvimento da tradução automática se iniciou nos primeiros anos da década de 1930 com as patentes introduzidas em 1933 por Georges Artsrouni, na França, e por Petr Trojanskij, na Rússia, para o desenvolvimento de um dispositivo que poderia ser utilizado como um dicionário de tradução multilíngue (Kenny, 2018: 429).

[3] Para uma descrição mais completa desses tipos de paradigmas, consultar o Anexo 1.

No memorando feito por Warren Weaver em 1949 para a *Rockefeller Foundation*, a partir da observação de "certain invariant properties which are, again not precisely but to some statistically useful degree, common to all languages" (Weaver, 1949: 2), esse pesquisador defende a existência de uma língua universal ainda não descoberta. Em seu memorando, ele sugere a observação do funcionamento lógico da linguagem e do contexto para a resolução de ambiguidades, bem como o uso de estatística e métodos criptográficos para auxiliar no desenvolvimento da tradução automática. Segundo Kenny (2018: 430) foi a partir do estímulo proporcionado por este memorando que uma onda de investimento em pesquisas em tradução automática surgiu nos EUA e no resto do mundo.

A primeira demonstração pública de um sistema em funcionamento foi feita em 1954, na Universidade de Georgetown. O pesquisador Léon Dostert, em colaboração com a *International Business Machines Corporation* (IBM), demonstrou a tradução de uma amostra de 49 frases do russo para o inglês. Apesar do reconhecimento da existência de limitações no sistema, a notícia foi recebida com bastante entusiasmo, levando ao aumento de grupos de pesquisa em tradução automática nos anos seguintes nos EUA, no Reino Unido e na Rússia, além de outros países. De acordo com Hutchins (2010: 2-3), entre as décadas de 1950 e 1960, até o aparecimento dos paradigmas baseados em *corpora*, os sistemas dessa época se dividiam basicamente em três tipos: tradução direta, interlíngua e *transfer*.

A partir de 1960, após essa onda de entusiasmo, se iniciou um processo de desilusão acerca da tradução automática devido a dificuldade em resolver problemas, como o dos "múltiplos significados" em tradução e a ambiguidade. Segundo Hutchins (2010: 5), essa desilusão culminou no levantamento de Yehoshua Bar-Hillel (1960), em que ele criticou fortemente a ideia da produção de *Fully Automatic High Quality Translation* (FAHQT) com resultados semelhantes aos dos tradutores humanos. De acordo com Kenny (2018: 431), devido aos limites dos computadores da época, que não tinham acesso a um conhecimento enciclopédico tão abrangente ou que possuíam uma menor capacidade de processamento, Bar-Hillel conclui que traduções de baixa qualidade completamente automáticas ou traduções de alta qualidade parcialmente automáticas seriam objetivos mais realistas.

Em 1964, os patrocinadores governamentais de tradução automática nos EUA solicitaram ao *National Science Foundation* a organização do *Automatic Language Processing Advisory Committee* (ALPAC) com o objetivo de examinar o cenário da tradução automática naquele momento. A onda de desilusão atingiu seu ápice após o

surgimento do relatório desse comitê em 1966. A ALPAC, após analisar os resultados das pesquisas feitas na área de tradução automática nos últimos anos, concluiu que a tradução feita pela máquina era mais lenta, menos fiável e duas vezes mais cara do que a tradução humana e que "while we have machine-aided translation of general scientific text, we do not have useful machine translation. Further, there is no immediate or predictable prospect of useful machine translation" (ALPAC, 1966: 32). A partir das conclusões desse relatório, houve uma imensa diminuição no financiamento de pesquisas nessa área no mundo inteiro, principalmente nos EUA, durante mais de uma década.

No entanto, mesmo após o surgimento do relatório da ALPAC, o interesse em tradução automática continuou a crescer em outras regiões por motivos políticos, como ressalta Kenny (2018: 432). No Canadá, a política bicultural do seu governo criou a crescente procura por traduções de textos entre o Inglês e o Francês e na Europa, a criação das Comunidades Europeias (atual União Europeia) resultou na demanda crescente de traduções de textos nessa comunidade multilíngue (Hutchins, 2010: 6).

No Canadá, as pesquisas começaram em 1970 com a procura de um *transfer* sintático para a tradução inglês-francês. O sistema Météo foi um dos resultados mais prósperos do projeto TAUM (*Traduction Automatique de l'Université de Montréal*). Esse sistema, designado especificamente para o vocabulário restrito e a sintaxe limitada dos relatórios de metereologia, foi apontado por Kenny (2018: 432), do ponto de vista da longevidade, como a implementação de tradução automática mais bem sucedida do século XX, tendo sido utilizado entre os anos 1977 e 2002.

O SYSTRAN foi um dos mais importantes sistemas do período posterior ao relatório da ALPAC. Esse sistema, desenvolvido por Peter Toma em 1968, foi utilizado por diversas organizações como a NATO e as Comunidades Europeias e por grandes empresas como General Motors, Dornier e Xerox (Hutchins, 2010: 7). Os maiores rivais desse sistema foram o Logos, utilizado principalmente no início dos anos 1980, e o METAL, que surgiu a partir do fim dessa década. A princípio, esses três sistemas foram designados para ter uma aplicação generalizada, apesar de na prática terem seus dicionários adaptados para domínios particulares.

Na Europa, a Comissão das Comunidades Europeias (CEC) começou a usar o sistema SYSTRAN em 1976, inicialmente para traduzir textos entre inglês e francês e posteriormente com a adição de novos pares linguísticos. Somente em 2010 o uso desse sistema foi descontinuado, tendo sido substituído por um sistema desenvolvido pela própria organização (Kenny, 2018: 432). O projeto Eurotra, desenvolvido no seio das Comunidades Europeias, foi apontado por Hutchins (2010: 8) como um dos melhores

projetos da década de 1980: seu objetivo era a construção de um sistema *transfer* multilíngue para tradução de textos através de todas as línguas participantes dessa comunidade. Apesar de o projeto ter sido terminado no início da década de 1990, ele conseguiu estimular a pesquisa em linguística computacional através da Europa.

A década 1980 foi marcado pela proliferação de sistemas comerciais, principalmente na América do Norte e no Japão, tendo alguns deles sido desenvolvidos para computadores pessoais, em crescimento nesta época. Kenny (2018: 432) aponta que, na maioria dos casos, pós-editores eram requisitados para melhorar a qualidade do texto produzido pela tradução automática. Houve também sistemas interativos como o ALPS *Translation Support System,* no qual eram feitas perguntas ao usuário para resolver problemas de ambiguidade durante o processo de tradução automática.

Apesar do constante aumento no consumo de traduções automáticas a partir dos anos 1990, os tradutores humanos profissionais se interessavam de maneira secundária nesse tipo de sistema. Kenny (2018: 434) explica que isso ocorreu devido à má qualidade das traduções produzidas pelos sistemas disponíveis na época e devido ao mercado de *computer aids for translations* ter sido significativamente desenvolvido e já usar memórias de tradução, tecnologias que permitiam a reutilização de traduções humanas.

Segundo Kenny (2018: 433), uma alternativa aos paradigmas de tradução que utilizavam regras linguísticas foi apresentada pela primeira vez por um grupo de investigadores da IBM, em 1988, e foi desenvolvida a partir dos anos 1990. O sistema Candide, desenvolvido pela IBM, produzia modelos probabilísticos de tradução a partir da *corpora* paralela de textos bilíngues do Parlamento Canadense. Os resultados das traduções desse sistema surpreenderam pesquisadores, pois o sistema produzia traduções aceitáveis sem utilizar regras linguísticas.

Apesar de ter sido inicialmente recebido com bastante incredulidade, no início dos anos 2000, os sistemas de *statistical machine translation* (SMT) ganharam importância (Kenny, 2018: 433). Eles foram auxiliados pela crescente disponibilidade de "bitextos" eletrônicos, pelo aumento progressivo de processamento computacional e armazenagem de memória, bem como pelo esforço colaborativo de desenvolvedores que disponibilizaram ferramentas *open-source* de SMT.

O segundo maior paradigma de tradução automática baseado em dados, conhecido como *example-based machine translation* (EBMT), começou a ser testado no fim da década de 1980, principalmente no Japão. Esse sistema envolvia a seleção e extração de frases equivalentes, a partir de um banco de dados de textos paralelos e o alinhamento

dessas frases, com base em métodos estatísticos ou regras linguísticas (Hutchins, 2010: 12).

A ampliação das técnicas e sistemas de tradução automática demonstrou as desvantagens em adotar um só paradigma para resolver os problemas de tradução. Tendo isso em vista, foram desenvolvidos paradigmas híbridos que combinavam técnicas de *rule-based*, *statistics-based* e *example-based* (Hutchins, 2010: 14).

Segundo Kenny, o crescimento da tradução automática a partir dos anos 1990 e principalmente nos anos 2000 foi favorecido pelo aumento do acesso à Internet e a outras tecnologias, o que causou o aumento da diversidade linguística encontrada online (2018: 433). Diversos desenvolvedores de tradução automática começaram a fornecer serviços de tradução automática online a partir dos anos 1990, como os sistemas Systran, Reverso, AltaVista e PARS (Hutchins, 2010: 17). A Google Translate é apontada por Kenny como um exemplo da onda de crescimento do uso e da disponibilidade de tradução automática a partir dos anos 2000 (Kenny, 2018: 433).

Os sistemas SMT foram difundidos de forma generalizada devido a sua boa performance. Por isso, foram considerados a forma mais avançada e eficiente de tradução automática até o lançamento do sistema NMT desenvolvido entre 2015/16, que surgiu como o paradigma de tradução automática mais prometedor dos últimos anos, demonstrando uma performance superior em *benchmarks* públicos e sendo adotado rapidamente por empresas como Google e SYSTRAN (Koehn e Knowles, 2017: 1). No fim de 2016, esse sistema já estava sendo utilizado pelas maiores empresas de tecnologia para fornecer tradução através de diversas plataformas como a Google Translate (Wu et al., 2017; Johnson et al., 2017) e a Microsoft. Outros utilizadores e fornecedores especializados também começaram a mudar para a nova tecnologia, auxiliados por ferramentas de NMT fornecidas gratuitamente (Kenny, 2018: 434).

Tendo em vista o recente lançamento do NMT e seu processo estar em desenvolvimento e aprimoramento, é possível encontra opiniões divergentes acerca dos resultados fornecidos pelo sistema. Para exemplificar, por um lado, Bentivogli et al. afirmam que “the outcomes of the analysis confirm that NMT has significantly pushed ahead the state of the art, especially in a language pair involving rich morphology prediction and significant word reordering” (2016: 9). Por outro lado, Castilho et al. sustentam que “while automatic evaluation results published for NMT are undeniably exciting, so far it would appear that NMT has not fully reached the quality of SMT, based on human evaluation” (2017: 118).

Quanto à qualidade da tradução, Bentivogli et al. verificaram que "NMT output contains less morphology errors, less lexical errors, and substantially less word order erros" (2016: 9). No entanto, Koehn e Knowles afirmam que "NMT systems have lower quality out of domain, to the point that they completely sacrifice adequacy for the sake of fluency" (2017: 28). Já Castilho et al. (2019) ressaltam que ainda nos encontramos nos primórdios do desenvolvimento dos sistemas NMT, por isso ainda é necessário fazer investigações mais aprofundadas com amostras maiores, envolvendo mais pares linguísticos e considerando-se diferentes níveis de experiência, como o uso dos resultados desse tipo de sistema em processos de pós-edição ou pré-edição.

Ao longo de seu desenvolvimento a tradução automática foi marcada por altos e baixos, sendo por vezes exageradamente elogiada, como no período após a primeira demonstração pública na década de 1950, e outras vezes excessivamente depreciada, como no relatório da ALPAC na década de 1960. Ainda hoje essa tendência pode ser observada no caso da NMT, como apontam Castilho et al. (2017: 1): "from the first commercial rule-based systems to more recent statistical models, there has, however, generally been great discrepancy between the high expectation of what MT should accomplish and what it is actually able to deliver".

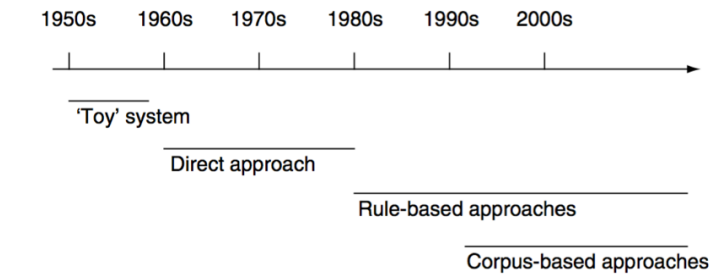

Figura 1 – Cronologia do desenvolvimento da Tradução Automática
(extraída de Quah, 2006: 58)

A figura 1 pode resumir, de forma generalizada, o desenvolvimento histórico da tradução automática a partir da segunda metade do século XX. É possível observar que apesar do surgimento dos paradigmas baseados em *corpora*, no início dos anos 1990, os sistemas baseados em regras não pararam completamente de ser utilizados e pesquisas continuam a ser feitas acerca de regras linguísticas que possam auxiliar o processo de tradução. O mesmo não ocorreu com os sistemas que utilizavam a tradução direta, devido a sua imensa simplicidade.

2.2 Paradigmas de Tradução Automática: o funcionamento do NMT

Os parágrafos seguintes se dedicam a fornecer uma visão geral acerca do funcionamento do sistema *Neural Machine Translation* (NMT), visto ser o sistema utilizado pela empresa Unbabel[4].

Entretanto, antes de iniciar a discussão acerca desse tipo de sistema, é interessante a diferença entre o tipo de arquitetura e o tipo de paradigma: "one does not pressuppose the other. The former refers to the actual processing design (i.e., direct, transfer, interlingual), whereas the later refers to informational componentes that aid the procesing design (knowledge-based, example based, statistics-based, etc.)" (Dorr et al., 1999: 18). Essa diferença também é ressaltada por Costa-Jussà et al. (2015: 4) que classificam os sistemas de tradução automática segundo dois critérios: o nível de representação e a fonte da informação.

Em relação ao nível de representação, há três arquiteturas possíveis: direto, *transfer* e interlíngua. Esses três processos podem ser resumidos no diagrama introduzido pelo pesquisador francês B. Vauquois em 1968:

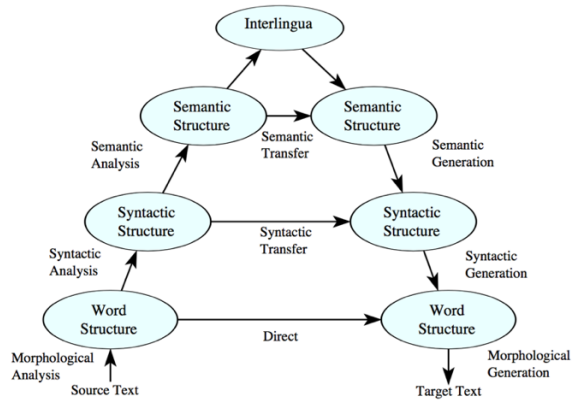

Figura 2 – Pirâmide de Vauquois

(extraída de Dorr et al., 1999: 13)

No método direto (*direct approach*), representado pela base da Pirâmide de Vauquois, há uma única transformação entre a língua de chegada (doravante LC) e a língua de partida (doravante LP), com uma análise superficial da LP e sem a geração da LC: as

[4] Considerando-se a extensão e os objetivos da presente pesquisa, um resumo acerca dos outros tipos de sistemas foi inserido no Anexo 1.

palavras são diretamente substituídas por palavras correspondentes na LC. O processo *transfer*, representado pelas seções intermediárias da pirâmide, constitui-se de três fases: análise, *transfer* e geração. Já no método de interlíngua, há a procura de uma representação universal de todas as línguas. Nesse processo, representada pelo topo da pirâmide, a LP é transformada na representação de interlíngua, através da análise, e depois essa representação é transformada no texto na LC. Nesse método, não há necessidade da etapa de *transfer*.

Segundo Dorr et al. (1999: 13), os três níveis da pirâmide correspondem aos diferentes níveis de *transfer*, dependendo da profundidade da análise fornecida pelo sistema: a mais superficial observa a estrutura sintática e morfológica da frase e a mais profunda observa os aspectos semânticos do texto. Na base do sistema há o método direto, que consiste na forma mais primitiva de *transfer* (substituição palavra por palavra). No topo da pirâmide está o método interlíngua que consiste na forma mais degenerada de *transfer*, ou seja, o mapeamento do *transfer* é praticamente inexistente. A maioria dos sistemas de tradução entram em algum lugar entre esses dois extremos, oscilando entre uma análise sintática superficial e uma análise semântica mais profunda.

Em relação à fonte da informação, os paradigmas de tradução automática podem ser divididos em dois grandes grupos: por um lado há os paradigmas que se baseiam em conhecimento linguístico, em que a tradução é feita a partir da produção de regras gramaticas e lexicais, denominados *rule based machine translation* (RBMT); por outro lado existem os paradigmas que se baseiam em *corpora*, denominados *corpus based machine translation* (CBMT), nesse caso a tradução é produzida a partir da comparação de textos já traduzidos por humanos e de cálculos estatísticos para descobrir qual tradução é a mais provável para uma data entrada. Estão inseridos nesse último tipo de paradigma os sistemas EBMT, SMT e NMT.

Assim como os SMT, os *neural machine translation* (NMT) aprendem a traduzir a partir *corpora* paralelos, mas fazem isso através de um método computacional diferente que utiliza redes de neurônios (Forcada, 2017: 293). Apesar dessa diferença, Koehn ressalta que NMT "is not a drastic step beyond what we have traditionally done in statistical machine translation (SMT)" (2016).

Esse recente sistema de tradução automática utiliza redes de inteligência artificial, nas quais milhares de unidades individuais, ou "neurônios" artificiais, análogos aos neurônios do cérebro humano, estão conectados entre si e a ativação de um neurônio depende do estímulo recebido do seu precedente e do "peso" da conexão na qual esse estímulo é propagado (Forcada, 2017: 293). É o estado de ativação de grandes grupos de neurônios

interconectados que pode ser entendido como representações das palavras individuais e dos contextos nos quais elas aparecem.

Segundo Kenny (2018: 436), treinar uma rede de neurônios para tradução automática significa basicamente aprender quais são os pesos que resultarão nas representações que podem melhor garantir que a rede, quando na fase de tradução (*decoding*), fornecerá resultados de traduções mais próximas possíveis do *gold standard* humano encontrado na fase de treinamento. Forcada (2017) dá mais detalhes acerca desse processo de treinamento:

> "We want the neural network to read each source sentence to form distributed representations (values of activations of groups of neurons), such that outputs computed from them are as close as possible to the corresponding reference or *gold-standard* translations in the training set (ideally produced by translation professionals). To that end, one trains the neural network; that is, determines the weight or strength of each of the connections between neurons so that the desired results are obtained" (Forcada, 2017: 294).

As representações não são construídas em uma só parte, mas sim em diversas camadas constituídas de listas com uma quantidade numérica fixa. Kenny (2018: 436) ressalta que somente as camadas externas, o *input* e o *output* da rede, podem ser analisadas pelos humanos, pois as camadas internas permanecem "escondidas".

Assim como os sistemas estatísticos, os sistemas neuronais são treinados a partir de dados e o aprendizado dos pesos usados em NMT é semelhante ao aprendizado das probabilidades de tradução em SMT. Porém, Kenny (2018: 436) ressalta que os sistemas NMT possuem uma arquitetura monolítica mais simples e processam sentenças inteiras ao invés de somente *n*-grams.

Bentivogli et al. (2016: 1) apontam o uso de "*attention mechanism*" como uma evolução que marcou o funcionamento dos NMT e o seu melhoramento. Como Kenny (2018), ele também ressalta que por um lado, os sistemas NMT representam uma simplificação em relação aos paradigmas anteriores, mas, por outro lado, o processo dos NMT é menos transparente (Bentivogli et al., 2016: 1).

Segundo Wu et al. (2016: 1), a arquitetura de uma NMT é tipicamente composta de dois *recurrent neural network* (RNN), um para consumir a sequência de texto do *input* e outro para gerar o *output* do texto traduzido. Esse sistema é frequentemente acompanhado por um *attention mechanism,* que auxilia a lidar com longas sequências de *input*. Para exemplificar o funcionamento da NMT, citamos a seguir os componentes da NMT desenvolvida pela Google Translate:

> "It has three components: an encoder network, a decoder network, and an attention network. The encoder transforms a source sentence into a list of vectors, one vector per input symbol. Given this list of vectors, the decoder produces one symbol at a time, until the special end-of-sentence symbol (EOS) is produced. The encoder and decoder are connected through an attention module which allows the decoder to focus on different regions of the source sentence during the course of decoding" (Wu et al., 2016: 3).

Para auxiliar na visualização desse processo, apresentamos uma figura apresentada por Luong et al. (2016):

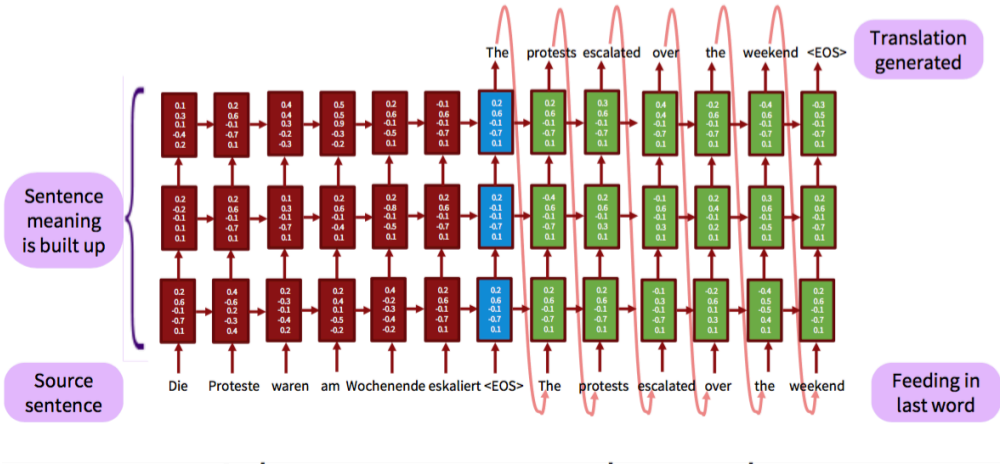

Figura 3 – Funcionamento do NMT

(extraída de Luong et al., 2016: 22)

Primeiramente, é feita representação interna da frase na LC a partir de uma sequência de vetores. Nessa etapa, cada palavra da frase na LP passa por um *encoder* e gera diferentes camadas de vetores, utilizando a palavra atual e o contexto processado anteriormente. Quando a sequência completa de vetores do texto de partida é produzida, uma por uma, as palavras na LC são geradas. Para isso, o *decoder* utiliza a lista de vetores do contexto de chegada gerado juntamente com as camadas ligadas à palavra anterior. Há também o *attention model* que auxilia na verificação do texto na LP. É através do uso de todos esses elementos que "the decoder provides, at each position of the target sentence being built, and for every possible word in the target vocabulary, the likelihood that the word is a continuation of what has already been produced" (Forcada, 2016: 294).

3. A UNBABEL

Neste capítulo serão retratadas as principais características do local em que decorreu o estágio no qual foram recolhidos os dados que deram origem à presente dissertação. Assim, nesta seção introdutória, serão apresentadas as principais características da empresa Unbabel; na seção 3.1, será explicitado o fluxo de trabalho da Unbabel e será esclarecido o funcionamento de ferramentas que auxiliam na identificação dos erros de tradução; na seção 3.2, serão abordadas sucintamente as *Language Guidelines*, orientações dadas aos pós-editores.

A Unbabel é uma empresa portuguesa fundada em 2013 que fornece serviços de tradução para outras empresas através da combinação de tradução automática e da revisão feita por editores humanos. Pode ser definida como uma *start up*, ou seja, uma empresa inovadora que acabou de ser lançada e está em rápido crescimento. Por isso, os processos e as ferramentas citados aqui podem ter sido ligeiramente modificados até a data de publicação da presente pesquisa. Estudos feitos anteriormente nessa *start up* demonstram essa mudança constante (cf. Comparin, 2016; Testa, 2018; e Figueira, 2018).

Quanto aos sistemas de tradução automática utilizados nessa empresa, em 2016 ela utilizava o sistema de tradução automática SMT fornecido pela Google (Comparin, 2016), em setembro do mesmo ano ela passou a utilizar a tradução automática SMT *open-source* do sistema Moses (Testa, 2018), durante o período do estágio referido na presente pesquisa, entre o segundo semestre de 2017 e o primeiro semestre de 2018, a Unbabel já tinha desenvolvido seu próprio sistema de NMT e o estava utilizando para produzir suas traduções (cf. Silva et al., 2018).

Tendo em vista as características dos sistemas NMT, Unbabel (2017) define a rede de tradução realizada na Unbabel como *self-learning*, pois os textos multilíngues traduzidos pelo sistema são reutilizados como *corpora* de treinamento para a máquina (Silva et al. 2018).

A Unbabel recebe os textos de partida de outras empresas diretamente através da API ou através de integrações com plataformas como Zendesk, Salesforce ou Freshdesk. Esses textos geralmente estão relacionados com serviços de atendimento ao cliente, como *tickets* (e-mails), *FAQ* (perguntas frequentes) e *chats* (comunicação escrita em tempo real). Eles também podem ser conteúdos de *blogs*, descrições de produtos, avaliação de clientes e outros conteúdos produzidos pelos usuários (*user-generated content*). Em vista disso, a Unbabel funciona como um agente intermediário entre a empresa que solicita

seus serviços de tradução e os clientes dessa mesma empresa. Os editores que revisam a tradução produzida pela máquina também são membros dessa equação e, durante a pós-edição, devem considerar não somente a Unbabel, mas também as instruções da empresa que enviou o texto de partida.

Na Unbabel, a pós-edição humana é feita por ampla comunidade de editores, através de 28 línguas, utilizando-se o princípio de *crowdtranslation*, um conceito inovador derivado da noção de *crowdsourcing*[5] (Unbabel, 2014). Nesse tipo de pós-edição, a revisão é feita por pessoas que não são tradutoras profissionais: a "*crowd*", que pode ser usada para criar traduções rapidamente através da divisão do texto em pequenas tarefas (Moorkens, 2017: 2). As inovações trazidas por esse tipo de tradução estão pouco a pouco afetando o rumo da tradução: "even though crowdsourcing is still a niche activity and affects the sector only to a limited extent, its influence is bound to grow and there are useful lessons on good practices to be learnt also for professional translators" (European Commission, 2012: 35).

Antes de começar a revisar, os editores são treinados e avaliados automaticamente a partir de *training tasks,* ou seja, tarefas que simulam o conteúdo que irão traduzir (Unbabel, 2014). Também tradutores profissionais ou doutorandos avaliam o trabalho dos editores e fornecem *feedback*. Os editores têm acesso a *guidelines* gerais de como utilizar a plataforma e a *guidelines* específicas relacionadas com aspectos linguísticos de cada uma das línguas traduzidas pela Unbabel.

3.1 Fluxo de trabalho

Para combinar o sistema de tradução automática, os pós-editores e as outras ferramentas utilizadas em seu processo de tradução, a Unbabel possui um fluxo de trabalho, denominado *Unbabel's Translation Pipeline*, que consiste numa sequência de passos entre o texto de chegada recebido pela Unbabel e o texto de partida enviado aos seus clientes.

Antes de o texto a ser traduzido entrar nesse fluxo, os glossários e as instruções dos clientes são elaborados pelas equipes da empresa. O glossário consiste em traduções preferenciais para termos importantes do conteúdo dos textos, garantindo um mínimo de consistência textual (Unbabel, 2019a). As instruções do cliente podem fornecer

[5] "Crowdsourcing is the process by means of which organisations can tap into the wisdom of their dedicated external community and use the wisdom for their benefit, i.e. with low cost, for more languages, and within the specified time frame" (Anastasiou e Gupta, 2011: 2).

informações gerais acerca da empresa relacionada com o texto e indicações acerca do registro (formal ou informal) a ser usado.

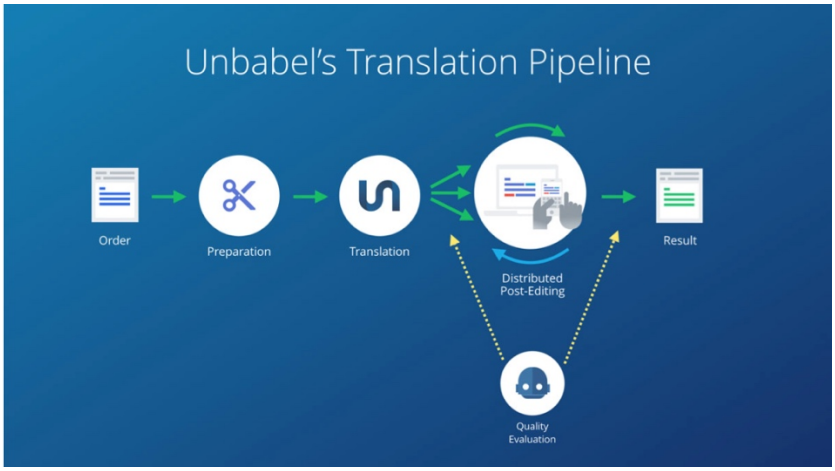

Figura 4 – Fluxo de Trabalho da Unbabel
(extraída do site da Unbabel, 2018)

Para simplificar a elucidação da *Unbabel's Translation Pipeline*, esse fluxo de trabalho pode ser dividido em quatro passos principais, ilustrados na figura 4: preparação, tradução automática, avaliação automática da qualidade e pós-edição humana. Esses passos serão vistos brevemente nos parágrafos seguintes.

Na fase de preparação, o texto a ser traduzido é analisado e certos fatores que influenciarão seu caminho ao longo da *pipeline* são detectados. É nessa fase que os glossários customizados e as instruções do cliente são assinalados no texto de partida. Além disso, através do processo de anonimização, toda informação considerada sensível e todo dado possível de levar à identificação pessoal é retirado.

Após a preparação, o texto é enviado para o sistema de tradução automática, que inicialmente verifica e utiliza suas "memórias de tradução" (Unbabel, 2018). Um modelo de tradução é então construído a partir da junção dos dados dessa análise com outras informações como o registro e o tópico do texto.

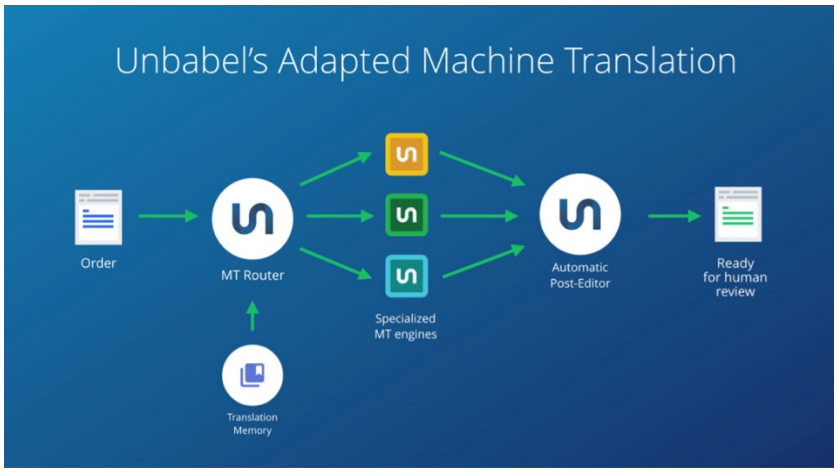

Figura 5 – Tradução Automática na Unbabel

(extraída do site da Unbabel, 2018)

A próxima etapa, representada na figura 5, ocorre no *Machine Translation Router,* que escolhe as melhores máquinas especializadas de tradução automática, baseando-se no conteúdo, na área envolvida e no cliente. Depois disso, o texto passa pelo *Automatic Post-Editing* (APE) através do qual a Unbabel pode melhorar a qualidade das traduções mediante uma pós-edição automática dos erros identificados pelo APE (Unbabel, 2017c).

Essa versão revisada automaticamente é então avaliada pelo sistema de avaliação automática *Open Kiwi* desenvolvido pela Unbabel (Martins et al., 2019). A qualidade é verificada pela máquina em duas etapas do processo: após a tradução feita pelo sistema e após a revisão feita pelos editores. Antes de enviar o texto aos editores, as partes do texto consideradas incorretas pelo sistema de QE são sublinhadas, para que seja efetuada a correção.

Na fase de pós-edição humana, as partes do texto traduzido são enviadas para a comunidade de editores através da plataforma de edição da Unbabel. Os pós-editores têm acesso ao texto original na LP e ao texto traduzido pela máquina na LC. Eles devem corrigir os erros encontrados no texto, seguindo as orientações das *Language Guidelines* (definidas pela equipe de Qualidade da Unbabel, composta de linguistas especializados) e as instruções do cliente. Os pós-editores são auxiliados pelas seguintes ferramentas disponíveis na plataforma: o *Smartcheck*, o contexto de tradução, o glossário, as memórias de tradução e as instruções do cliente.

As memórias de tradução são fragmentos do texto que fazem parte do banco de dados do sistema da Unbabel e consistem em traduções válidas já utilizadas outras vezes. O contexto do texto a ser traduzido também pode ser fornecido na plataforma para auxiliar

em questões de gênero e número, por exemplo. Os dois parágrafos iniciais no texto de partida, apresentado na figura 6 a seguir, ilustram esse tipo de contexto:

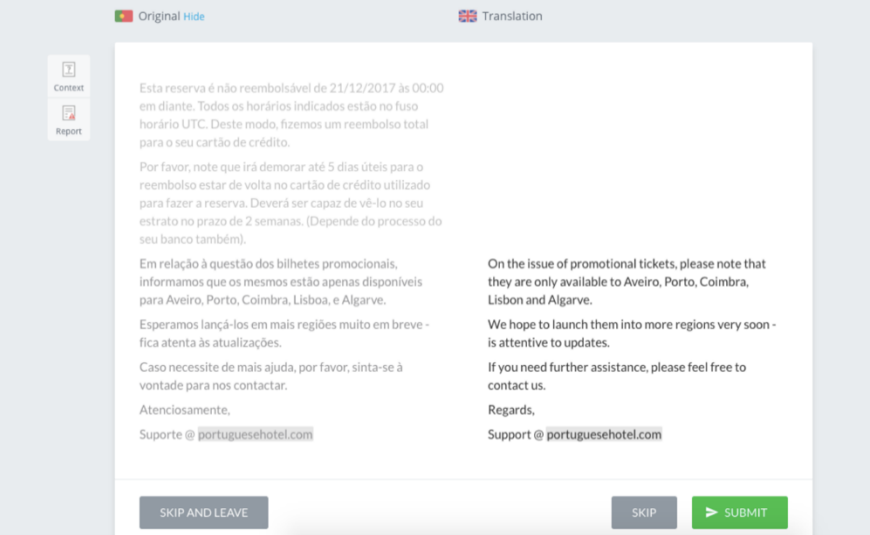

Figura 6 – Contexto na ferramenta de pós-edição
(extraída do site da Unbabel, 2019a)

O *Smartcheck* usado pela Unbabel é uma versão sobrecarregada da correção gramatical geralmente encontrada em editores de documentos (cf. figura 7). Ele verifica uma gama de potenciais erros e insere sugestões relacionadas com a ortografia, o registro, a coerência lexical (concordância sujeito-verbo, correspondência entre pronomes, gênero, etc.) e outras regras associadas às exigências do cliente. Através dessa ferramenta é possível acelerar a correção do texto.

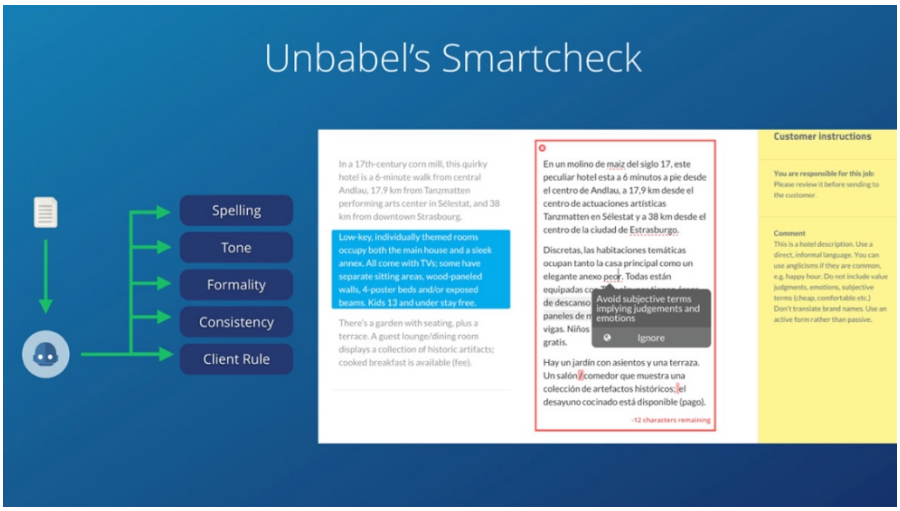

Figura 7 – *Smartcheck*

(extraída da Unbabel, 2018)

Segundo Unbabel (2017c), essa ferramenta foi desenvolvida com o auxílio de pesquisadores em *Natural Language Processing* e linguistas computacionais. Na prática, através de um amplo conjunto de regras, essa ferramenta destaca uma palavra ou um grupo de palavras com a cor verde, que indica sugestões dadas pelo sistema, ou vermelha, que indica erros críticos que devem ser corrigidos antes do envio da tarefa. Ao selecionar a palavra sublinhada, o *Smartcheck* fornecerá uma lista de palavras sugeridas, que podem ser aceitas através de um clique.

É importante ressaltar o funcionamento do *dependency parser,* visto o seu papel crucial na detecção dos erros no texto de chegada. O *parser* é uma importante ferramenta no processo de tradução automática, tendo em vista a possibilidade de indicar as relações de dependência sintáticas entre os elementos da frase e fornecer uma *tagging* (identificação) das categorias gramaticais de cada palavra. Essas informações podem auxiliar na resolução das ambiguidades sintáticas e semânticas encontradas na frase após a tradução, pois para cada palavra é indicada a sua categoria gramatical e valores para certas características como gênero, número, pessoa, modo, tempo, forma verbal, etc. Também é fornecida uma árvore de dependência representando a estrutura sintática da frase (Testa, 2018).

O *Turbo Parser* utilizado pela Unbabel foi desenvolvido por Martins et al. (2013): "We presented new third-order non-projective parsers which are both fast and accurate. (...) Results are above the state of the art for large datasets and non-projective languages" (2013: 5). Esse *parser* auxilia na análise dos dados e fornece informações mais específicas

ao *Smartcheck* acerca das estruturas com o objetivo de melhorar as sugestões feitas por essa ferramenta. O *Turbo Parser* não é visível diretamente pelos editores, mas está integrado na análise feita pelo *Smartcheck*.

Finalmente, após a fase de pós-edição, o texto é enviado para mais uma avaliação automática da qualidade e, caso seja considerado adequado, as partes do texto são aglutinadas em um só texto e esse último é enviado ao cliente.

3.2 *Language Guidelines*

A partir da análise dos principais erros feitos pelos editores e pelo sistema de tradução automática, a Unbabel elaborou as *Language Guidelines*, que consistem em orientações específicas para cada uma das línguas traduzidas pela Unbabel. Essas *Guidelines* fornecem aos pós-editores instruções gerais acerca do funcionamento da plataforma e instruções específicas acerca das particularidades e dos erros mais comuns na tradução de cada uma dessas línguas.

Interessa-nos as *Language Guidelines* do PB, principalmente nos segmentos referentes aos erros mais frequentes, pouca edição e erros no texto de partida (Unbabel, 2019a). Essas *Guidelines* fornecem instruções gerais sobre a edição do texto traduzido, como, por exemplo, evitar pouca edição, fazer o texto soar natural e fluído e corrigir erros encontrados no texto de partida: "faça o melhor possível para ajustar a sua tradução ao público falante nativo da língua de chegada" (Unbabel, 2019a).

O segmento das *Guidelines* que trata dos "erros mais frequentes" é o mais informativo para a presente pesquisa, pois oferece instruções específicas para o PB a nível morfológico, sintático e semântico. Esse segmento discute acerca de erros específicos de tradução entre inglês-PB com exemplos e sugestões para a correção dos seguintes tipos de erros: exatidão, fluência, estilo, terminologia, variação incorreta na mesma língua e nomes de entidades (cf. Tabela 2.1 no Anexo 2). Tendo em vista os objetivos da presente pesquisa, as porções mais relevantes para a análise são nomeadamente "Tradução literal", "Ordem das palavras" e "Variação incorreta da mesma língua", por isso as informações contidas nessas porções foram retomadas na seção 6 (cf. Tabela 2.2 no Anexo 2).

4. PROCESSO DE ANOTAÇÃO

Neste capítulo será apresentado o processo de anotação utilizado pela Unbabel. Nesta seção introdutória, será feita uma síntese geral acerca do desenvolvimento da avaliação da tradução automática.

Na seção 4.1, será abordado o desenvolvimento da tipologia de erros utilizada pela Unbabel, que se baseia na *Multidimentional Quality Metrics* (MQM).

Na seção 4.2, a ferramenta de anotação de erros desenvolvida pela Unbabel será apresentada em quatro subseções: 4.2.1 aborda o funcionamento geral da plataforma; 4.2.2 trata do processo de segmentação dos erros; 4.2.3 apresenta a tipologia de erros utilizada na Unbabel; e 4.2.4 apresenta os níveis de severidade.

Devido à própria natureza das línguas naturais, que são heterogêneas e não exprimem o mesmo conteúdo da mesma maneira, a avaliação da tradução é uma tarefa complexa (cf. Han, 2016). A resposta "correta" não está clara quando se trata de línguas, principalmente no que envolve a tradução, por isso o julgamento sobre o que está ou não errado no resultado da MT ainda possui um grau de subjetividade (Dorr, 1999: 36). Para Lommel (2015), "a quality translation demonstrates required accuracy and fluency for the audience and purpose and complies with all other negotiated specifications, taking into account end-user needs" (2014: 2-3). Por conseguinte, a qualidade de uma tradução depende primordialmente do nível de fiabilidade e de fluência, bem como das especificações requisitadas pelo usuário dessa tradução.

A avaliação do resultado final da tradução pode ser feita através de processos automáticos e/ou humanos, sendo que cada um deles tem suas vantagens e desvantagens. Segundo Gornostay (2008: 4), as métricas humanas, objeto de investigação da presente pesquisa, envolvem o verdadeiro fim do texto traduzido, ou seja, o ponto de vista humano, e possibilitam uma análise mais profunda do processo de tradução; contudo, a avaliação humana é mais cara e demorada, avaliadores bilíngues treinados são necessários, e não há um método de comparação adequada entre os variados sistemas.

Como apontam Lommel et al. (2014: 31), o desenvolvimento dos sistemas de Tradução Automática faz grande uso do conhecimento e julgamento humano acerca da qualidade da tradução. Na Unbabel, essa utilização do conhecimento humano para a melhoria de seus sistemas pode ser vista no processo de pós-edição, já abordado na seção 3.1, e também no processo de avaliação humana da qualidade, que será tratado na presente seção.

Como assinalam Lommel et al., (2014), uma das principais formas de verificar a validade e fiabilidade de um método de avaliação é observar se os anotadores são capazes de aplicar esse método de maneira consistente e uniforme. Ao analisar os dados da anotação humana do projeto QTLaunchPad, esses pesquisadores verificaram que "inter-annotator agreement (IAA) is relatively low, in part because humans differ in their understanding of quality problems, their causes, and the ways to fix them" (Lommel et al., 2014: 31).

Segundo esses autores, além de outros fatores, os anotadores podem discordar a respeito da existência ou ausência de determinado erro, acerca da definição do intervalo (*span*) em que se encontra o erro, bem como da sua categorização e do seu grau de severidade. A partir desses problemas, Lommel et al. (2014) modificaram as instruções dadas aos anotadores, incluindo instrumentos que podem auxiliar no processo de tomada de decisão. Contudo, eles ressaltam que, apesar dessas melhorias, já era possível prever que os problemas que causam em discordância entre os anotadores permanecerão e são inerentes à tarefa de avaliação da qualidade, pois segundo eles, "to a large extent this disagreement reflects the variability of human language" (Lommel et al., 2014: 36).

4.1 Desenvolvimento da tipologia de erros usada pela Unbabel

Na busca de uma forma objetiva de determinar a qualidade da tradução, no fim da década de 1990 e no início do século XXI, houve o aumento do uso de listas "objetivas" de tipos de erros, exemplificadas pelo LISA QA Model, lançado *pela Localization Industry Association* (LISA) e baseado nas técnicas desse grupo fornecedor de serviços de localização, e pelo padrão desenvolvido pela *Society of Automobile Engineer International* (SAE), para a avaliação da qualidade dos manuais de serviços automotivos (Lommel et al., 2014: 457).

Segundo Lommel et al. (2014), esses modelos compartilhavam algumas características comuns: ambos forneciam uma lista de erros (25 para o LISA QA Model e 7 para o SAE J2450) que poderiam ser relacionados com erros específicos do texto e os dois também classificavam os erros a partir de graus de severidade. Esses dois modelos foram difundidos de maneira generalizada, sendo frequentemente customizados para corresponder às exigências específicas dos usuários.

Além desses modelos, Lommel et al. (2014) apontam que outras ferramentas de verificação automática da qualidade foram desenvolvidas durante esse período. As

métricas de avaliação automática como *Bilingual Evaluation Understudy* (BLEU; Papineni et al. 2002) e *Metric for Evaluation of Translation with Explicit ORdering* (METEOR; Banerjee e Lavie, 2005), por exemplo, baseavam-se em princípios diferentes dos que são aplicados aos instrumentos de avaliação humana. Por isso, os resultados fornecidos por sistemas automáticos e humanos têm sido geralmente incomparáveis.

A métrica *Multidimensional Quality Metrics* (MQM) foi desenvolvida pelo projeto QTLaunchPad, financiado pela União Europeia, para lidar com as limitações das avaliações de qualidade anteriores. Essa métrica se originou a partir dos esforços de atualização do LISA QA Model e consequentemente possui princípios semelhantes (Lommel et al., 2014: 458). Entretanto, diferente dos modelos de avaliação de qualidade anteriores, desde o início a MQM foi projetada para ser flexível e trabalhar com outras normas, buscando uma avaliação de qualidade que possa ser integrada no ciclo de produção do documento (Lommel et al., 2014: 459).

Dado o seu multilinguismo característico e a necessidade crescente de traduções de textos em todas as suas 27 línguas oficiais, a União Europeia (UE) procura constantemente desenvolver as capacidades da tradução automática. Foi a partir dessa necessidade que surgiu o projeto de tradução automática *Quality Translation 21* (QT21), financiado pelo *Horizon 2020*, programa de investigação e inovação da UE. Esse projeto visa a eliminação de barreiras entre os países da UE, incluindo barreiras linguísticas, para a livre circulações de pessoas, informações e serviços (QT21 2019a). A partir dos objetivos desse projeto, que visam a melhoria da avaliação da qualidade, foi desenvolvida a MQM da QT21, estrutura para a análise de erros de tradução automática. O padrão industrial da TAUS, DQF (*Dynamic Quality Framework*), harmonizou-se com a MQM posteriormente.

Como Lommel et al. (2014: 456) ressaltam, a MQM não visa criar um modelo de avaliação *one-size-fits-all*: o objetivo desse modelo é a elaboração de um sistema aberto, extensível e flexível para a designar e descrever métricas da qualidade da tradução através de um vocabulário compartilhado de tipos de erros. Ela fornece um vocabulário de tipos de erros com nomes padronizados e definições para cada uma das categorias de erro usadas para analisar aspectos da qualidade da tradução e para identificar problemas que precisam ser resolvidos. A partir dessas definições, a MQM permite que os avaliadores saibam exatamente a qual tipo de erro a categoria se aplica. Essa uniformidade também permite que os resultados das avaliações sejam comparáveis. A referida métrica pode ser aplicada a traduções feitas através de processos humanos ou automáticos e a qualquer tipo de texto (Lommel et al., 2014: 456).

Essa tipologia de erros fornecida pela MQM pode se sustentar em diversos níveis de sofisticação e segmentação, sendo possível aplicar métricas simples com somente duas categorias, até métricas mais complexas com múltiplas categorias, segundo as necessidades específicas do usuário. Na MQM, os tipos de erro são organizados segundo uma hierarquia como a apresentada na MQM *Core*:

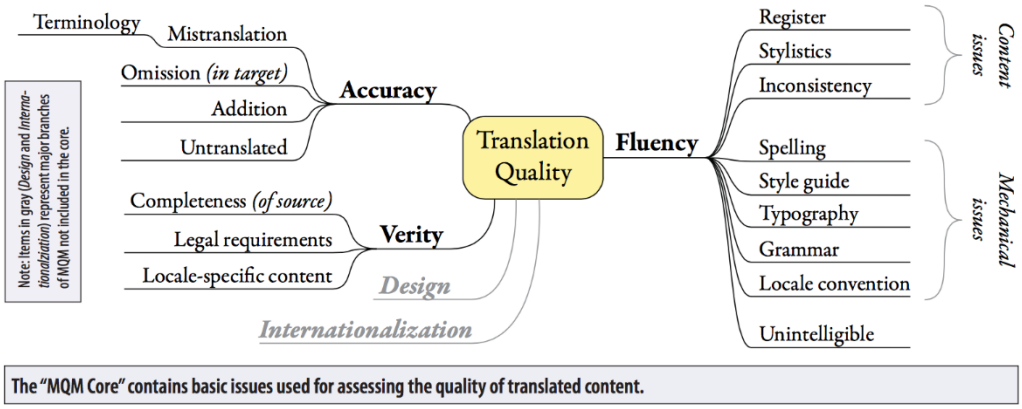

Figura 8 – MQM *Core*

(extraída de QT21, 2014)

Como é possível verificar na figura 8, o nível mais alto da tipologia de erros é composto por três categorias: *Accuracy*, *Fluency* e *Verity*. Os erros da primeira categoria se incluem na relação entre o texto de partida e o texto de chegada; os erros da segunda categoria abrangem o texto de chegada em si mesmo; finalmente, na terceira categoria é observada a relação entre o texto e o seu contexto exterior. As categorias *Design* e *Internalization* não são incluídas na MQM *Core*, mas também podem ser consideradas categorias primárias, dependendo das necessidades do utilizador (QTLaunchpad 2014b). A partir da base fornecida pela MQM, o utilizador pode customizar as tipologias de erros de acordo com a profundidade de análise desejada e segundo as categorias mais interessantes para a avaliação (Lommel e Burchardt, 2014: 4).

Outra característica essencial da MQM é a definição da *severity,* que indica o grau de severidade do erro naquele contexto. Essa definição é importante, pois os erros mais severos têm mais impacto na qualidade final do texto. Os quatro níveis de severidade básica da MQM são *neutral, minor*, *major* e *critical* (Lommel et al., 2014: 3). A MQM também auxilia na elaboração de um mecanismo de cálculo da nota final da qualidade, a partir da definição de pesos para cada tipo de erro, segundo a sua relevância e a influência desse erro na qualidade da tradução (Lommel et al., 2014: 11).

Quanto à customização das categorias de erros, por um lado, apesar da MQM permitir a subdivisão das categorias de erros em diversos componentes, no caso da avaliação humana é recomendado que as categorias não sejam excessivamente complexas, tendo em vista ser mais difícil para um avaliador humano fazer a distinção entre as subcategorias (Lommel et al., 2014: 459). Por outro lado, os utilizadores que participam de projetos de pesquisa e utilizam os dados resultantes das avaliações de qualidade, para verificar a origem do erro e aprimorar o processo de tradução, necessitam de uma taxonomia de erros mais complexa (Lommel e Burchardt, 2014: 5).

Além dessas recomendações, Lommel e Burchardt afirmam que as ferramentas de anotação *span-level*, ou seja, que permitem a conexão entre a categoria do erro e um intervalo determinado do texto, exigem mais treinamento dos avaliadores humanos, mas permitem uma visão mais detalhada acerca do erro, o que é benéfico para utilizadores que analisam mais profundamente os dados (Lommel e Burchardt, 2014: 6). Por fim, o treinamento dos avaliadores é assinalado por esses autores como essencial para a boa implementação da ferramenta de anotação, visto que métricas baseadas na MQM são geralmente complexas. Eles consideram que uma fase de treinamento, um documento com orientações (*guidelines*) e uma árvore de decisão (*decision tree*) são instrumentos que podem auxiliar nesse processo.

Quanto à taxonomia de erros utilizada pela empresa, Unbabel (2017c) revela que é utilizada a MQM "to be able to objectively compare our performance with third parties and open source translation libraries". Segundo as *Annotation Guidelines* dessa empresa, a identificação rigorosa e a etiquetagem dos erros, juntamente com um critério objetivo relativamente ao nível de severidade desses erros, é crucial para a avaliação da qualidade das traduções feita pela Unbabel (*Annotation Guidelines* 1.2).

A ferramenta de anotação desenvolvida por essa organização se caracteriza como *span-level* e possui uma tipologia de erros complexa. Como é possível verificar no site da Unbabel, a empresa se enquadra no grupo de utilizadores que investigam os erros de tradução para aperfeiçoar seu os seus métodos.

4.2 Ferramenta de anotação de erros na Unbabel

Como já foi explicitado na seção 3.1, a avaliação da qualidade da tradução produzida pela Unbabel é feita a partir de processos automáticos (Martins et al., 2017; Martins, 2019) e de processos humanos através da plataforma *Annotate*, também desenvolvida pela empresa (Unbabel, 2017c).

O principal objetivo do processo de anotação é a identificação dos erros presentes na tradução e a classificação deles segundo uma tipologia determinada. Ao encontrar um erro na tradução, o anotador humano deve aplicar uma etiqueta específica, segundo a tipologia de erros pré-estabelecida pela empresa, e escolher um nível de severidade, que correponde ao impacto que esse erro tem no sentido e na fluência geral da tradução (*Annotation Guidelines* 1.2).

Segundo Burchardt e Lommel (2014), a anotação é um processo que envolve uma grande exigência intelectual, por isso tradutores humanos experientes são considerados como anotadores ideais. Esses autores apontam que, mesmo após a fase de treinamento dos anotadores experientes, é possível encontrar variação entre as anotações. Por isso a importância de ter múltiplos anotadores para controlar a variabilidade entre os indivíduos.

Na Unbabel, a *Annotation* é feita por "a pool of specialists with backgrounds in Translation Studies and Linguistics, who are able to build a deep store of knowledge within our platform that boosts overall quality and decreases turnaround time to delivery" (Unbabel, 2017c). Antes de suas anotações serem consideradas como dados válidos para a avaliação da qualidade, os anotadores devem passar por um processo de treinamento, no qual eles são informados acerca do funcionamento geral da plataforma e recebem *Annotation Guidelines*[6], com instruções gerais e específicas acerca do processo.

Essas *Guidelines* são cruciais para o processo de anotação, pois elas fornecem diretivas a serem seguidas e elucidam eventuais dúvidas e dificuldades que os anotadores possam ter durante o processo. Essas orientações estabelecem parâmetros para que a seleção das categorias de erros e a avaliação da severidade sejam feitas de maneira uniforme, favorizando um maior *inter-annotator agreement*[7]. Segundo Burchardt e Lommel (2014: 11), as *Guidelines* são um dos materiais de treinamento mais úteis, mas também podem ser utilizadas como material de referência e, por isso, devem ser concisas e acessíveis.

Os dados provenientes das anotações humanas são consideravelmente relevantes para a análise do funcionamento geral das etapas de trabalho feitas na Unbabel. A partir dos resultados das anotações, grupos de linguistas e especialistas da empresa podem observar de maneira detalhada a qualidade da tradução automática e da pós-edição humana (Unbabel, 2017a).

---

[6] As *Annotation Guidelines* consideradas ao longo da presente pesquisa e apresentadas nesta seção correspondem à versão 1.2 vigente durante a anotação dos dados analisados, ou seja, entre novembro de 2017 e janeiro de 2018.

[7] "The inter-annotator agreement describes the degree of consensus and homogeneity in judgments among annotators" (Nowak e Ruger, 2011: 2).

As *Annotation Guidelines* da Unbabel ressaltam três objetivos específicos dos resultados da anotação humana: 1) treinar os modelos de tradução, pois os erros apontados são usados como um novo dado para o sistema, que, por sua vez, faz traduções mais apuradas; 2) alimentar os sistemas de Inteligência Artificial da Unbabel, que auxiliam os editores humanos na localização de potenciais erros; 3) auxiliar na execução de *Quality Audits*, ou seja, relatórios solicitados pelos clientes, nos quais são examinados os resultados dos serviços de tradução da Unbabel.

4.2.1 Funcionamento geral da plataforma *Annotate*

O processo de anotação é feito através de uma plataforma desenvolvida pela Unbabel: *Annotate Tool*. Ao selecionar uma das tarefas presentes na plataforma, a interface apresenta as seguintes áreas, ilustradas na figura 9: a barra superior, a barra de instruções do cliente, a área de texto da tarefa e o painel de anotação.

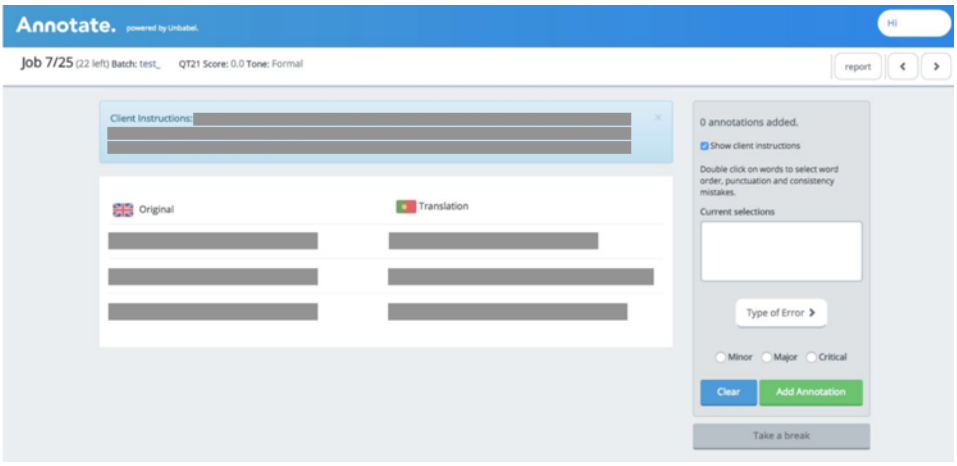

Figura 9 – Plataforma de anotação da Unbabel
(extraída de *Annotation Guidelines* 1.2)

No lado esquerdo do painel de anotação se encontra o painel de controle (figura 10) no qual é possível escolher a tipologia do erro, a partir de uma lista *drop-down,* e indicar a severidade (*minor*, *major* ou *critical*) para cada erro anotado. Além da tipologia de erros e da severidade, os anotadores podem adicionar um comentário e devem determinar a fluência do texto traduzido.

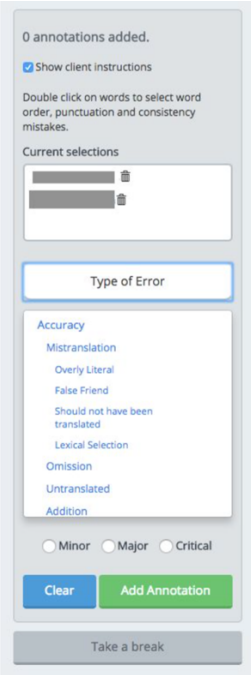

Figura 10 – Painel de controle: lista de tipologia de erros

(extraída de *Annotation Guidelines* 1.2)

Em geral, o processo de anotação consiste na identificação dos erros de tradução no texto de chegada, através da consulta dos glossários, do texto original e das instruções do cliente. Após essa etapa, o anotador deve selecionar o intervalo correspondente ao erro e escolher o tipo de erro a partir da lista fornecida pela plataforma. Cada categoria de erro se baseia em definições fornecidas pelas *Annotation Guidelines*. Em seguida, um dos três níveis de severidade deve ser atribuído para cada um dos erros indicados. Por fim, após a seleção de todos os erros, os anotadores devem obrigatoriamente apontar a fluência geral do texto numa escala de 1 a 5, em que 1 é a fluência mais baixa e 5 é a mais alta. Na Unbabel, essa escala de fluência se refere ao quão natural o texto traduzido soa na LC.

4.2.2 Processo de segmentação (*unitizing process*)

As *Annotation Guidelines* dedicam uma seção inteira para o esclarecimento do processo de segmentação dos erros, insistindo na sua importância e na ocorrência frequente de erros por parte dos anotadores. Esse processo é denominado *unitizing* pela empresa e consiste na seleção de um intervalo determinado, que pode incluir uma só palavra ou um conjunto de palavras, bem como espaços em branco, sinais de pontuação

ou símbolos. Cada seleção feita na ferramenta de anotação é considerada uma unidade (ou segmento).

Para orientar os anotadores, as *Guidelines* também fornecem exemplos de segmentação (*unitizing*) para os seguintes tipos de erros: *missing unit*, *wrong unit*, *word order issues* e *whitespace issues*. Ambas as categorias de erro *agreement* e *word order* são citadas nesses exemplos. Tendo em vista os objetivos do presente capítulo, a segmentação dessas categorias será vista com mais detalhes na seção 6.1.1.1 referente à análise desse processo.

4.2.3 Tipologia de erros nas *Annotation Guidelines*

A tipologia de erros utilizada pela Unbabel, baseada na MQM, foi adaptada às necessidades e aos tipos de textos traduzidos pela empresa. Os erros de tradução apresentados nessa taxonomia são representados por uma hierarquia composta de 3 troncos principais (etiquetas-mães) com o mesmo nível hierárquico: *Accuracy*, *Fluency* e *Style*[8], que correspondem a erros de fidelidade ao texto original, fluência e estilo, respectivamente.

A etiqueta *Accuracy* está relacionada com a relação entre o texto de chegada e o texto de partida: ela inclui erros em que o texto de chegada não representa corretamente o sentido do texto original. Para a etiqueta *Fluency* é necessário verificar a constituição do texto de chegada, independentemente de ser uma tradução ou não: ela inclui erros que afetam a leitura ou a compreensão do texto na LC. Os erros de *Style* se relacionam com o uso do registro, com a variante de língua utilizada e com as instruções fornecidas pelo cliente.

Essas etiquetas-mães incluem outras etiquetas mais específicas em níveis mais baixos de hierarquia: há um total de 17 etiquetas-filhas (6 em *Accuracy*, 6 em *Fluency* e 5 em *Style*); 25 etiquetas-netas (16 em *Accuracy* e 9 em *Fluency*); 8 etiquetas-bisnetas (todas em *Fluency*). Essas etiquetas podem ser consultadas nas Tabelas 3.1, 3.2 e 3.3 do Anexo 3. Quanto mais baixo é o nível hierárquico da tipologia, maior é a sua especificidade, pois etiquetas de nível mais baixo podem dar mais informações acerca do tipo de erro anotado do que as etiquetas mais altas. Na ferramenta da Unbabel, somente as etiquetas presentes nas "extremidades" dos troncos podem ser selecionadas, ou seja, a etiqueta de nível mais baixo deve ser selecionada.

[8] Por motivos de coerência e preservação do original, as designações dos erros citadas ao longo do relatório serão apresentadas na sua escrita original, em inglês, mesmo se existem correspondentes em português.

Além de definições acerca de cada tipologia, as *Guidelines* apresentam quadros com exemplos de cada uma das etiquetas usadas pela Unbabel, para resolver possíveis dúvidas e dificuldades no momento de escolher uma das tipologias. Além disso, a última seção das *Guidelines*, *Tricky Cases*, dedica-se à elucidação de casos mais difíceis de anotar. Nessa seção é possível encontrar indicações para a anotação de erros que envolvem mais de uma palavra, hífen, clíticos, palavras contraídas, entre outros erros. Essas orientações buscam uniformizar a anotação feita pelos anotadores, visando um maior *inter-annotator agreement*.

Mesmo com a disponibilização de exemplos para cada categoria de erro e com uma seção especial dedicada aos casos de anotação mais difíceis, ainda assim é possível encontrar divergências (*disagreements*) entre as anotações feitas pelos avaliadores da Unbabel. Algumas pesquisas feitas anteriormente nessa empresa apontam essa dificuldade. Comparin (2016: 41) ressalta que "there are cases in which an error can be categorized in different ways and the selection of the error type depends on the annotator's decision". Esse autora considera o processo de escolha das categorias de erros desafiador e faz uso de *decision trees* para auxiliar no treinamento dos anotadores de seu experimento. Como aponta Figueira (2018: 13-14), as etiquetas da Unbabel podem ser divididas em três grupos: não-ambíguas, ambíguas ou aparentemente ambíguas. Visto que em alguns casos as *Guidelines* não são claras acerca de qual categoria escolher, ele desenvolve uma *decision tree* para auxiliar na escolha das etiquetas de erros. Mais detelhes acerca dessas divergências entre as anotações serão abordados na seção 6.1, referente à análise dos dados quanto ao processo de anotação.

4.2.4 Severidade

Como já foi citado nas seções anteriores, na Unbabel os erros podem ser classificados em três diferentes níveis de severidade: *minor*, *major* e *critical*. A escolha da severidade do erro é crucial para a anotação, pois ela está diretamente relacionada com o resultado ou pontuação de qualidade do texto. As *Guidelines* ressaltam que a classificação dos erros não é rígida e depende muitos fatores, como a língua do texto anotado e o conteúdo presente no texto.

Segundo as *Annotation Guidelines*, a inserção do erro em uma dessas categorias depende do grau de influência que ele exerce na fidelidade ao texto original, na fluência do texto na LC, bem como na satisfação dos requisitos do cliente: os erros *minor* não levam a uma perda de significado e não confundem ou desorientam os leitores, mas podem diminuir a fluência e a qualidade estilística do texto de chegada. No caso do *major,*

os erros induzem os leitores ao erro, modificam o significado do texto (resultando no uso impróprio de um produto ou de um serviço) ou ocorrem numa parte importante do texto. Os erros de severidade *critical* modificam o significado original do texto e podem causar um impacto negativo à imagem do remetente da mensagem ou levar a implicações legais, financeiras ou ligadas à segurança e à saúde.

5. *DESCRIÇÃO DOS FENÔMENOS EM ANÁLISE*

Neste capítulo será apresentada uma descrição geral acerca dos fenômenos de concordância e ordem de palavras em PB para auxiliar na análise dos dados e na elaboração de sugestões para o processo de anotação e pós-edição. Antes de iniciar essa descrição, será feito um enquadramento histórico e social geral que têm influência no funcionamento desses fenômenos em PB.

Na seção 5.1, será fornecida uma visão geral da concordância em PB, tendo em vista os dados fornecidos pela empresa. Essa seção foi subdividida em três partes, abordando a concordância: no interior do SN (5.1.1); envolvendo o sujeito (5.1.2); e na coesão textual (5.1.3).

Na seção 5.2, serão apontadas as principais características da ordem de palavras em PB. Esta seção está subdividida em duas partes, tratando da ordem: de constituintes da frase (5.2.1), focando-se especialmente na ordem dos advérbios e dos clíticos, tendo em vista as suas particularidades; e de palavras internamente ao SN (5.2.2).

O contexto de variação do PB dificulta a elaboração de uma descrição uniforme dos fenômenos de concordância e ordem de palavras. Essa variação ocorre devido ao contato entre o português e as diversas línguas dos índios nativos e escravos africanos, desde o período da colonização no Brasil no século XVI e é perpetuada ainda hoje nas diferenças entre o português que é ensinado na escola e aquele falado pelos brasileiros. De acordo com Lucchesi, há "um generalizado sentimento de insegurança linguística entre os brasileiros, já que, por uma espécie de herança colonial, o padrão de correção gramatical adotado no Brasil é fortemente influenciado pelos modelos da língua de Portugal" (2012: 46-47).

Após a independência, no século XIX, a elite brasileira preferiu adotar um modelo baseado no português utilizado pela ex-metrópole, ao invés de abraçar as mudanças linguísticas que ocorreram no PB, ampliando as diferenças entre a língua falada por brasileiros menos escolarizados e a língua escrita (Lucchesi, 2012: 52). Esse modelo fundamentado nas regras do português europeu (doravante PE) ainda é empregado nos dias atuais por gramáticas tradicionais utilizadas em escolas brasileiras, como, por exemplo, as gramáticas de Bechara (1961/2002) e de Cunha e Cintra (1985), que por vezes apresentam descrições de estruturas baseadas no PE ou em textos antigos, desconsiderando as mudanças linguísticas pelas quais já passou o português padrão escrito no Brasil.

Em sua pesquisa acerca do ensino de gramática no Brasil, Duarte e Serra (2015) também apontam esse aspecto ao mencionar a enorme distância entre a fala e a escrita no PB e o fato de as gramáticas normativas e os livros didáticos apresentarem normas elaboradas "à luz das normas lusitanas de fins do Século XIX, que àquela altura já estavam bem distantes da escrita que vinha se desenvolvendo no Brasil desde o período colonial" (Duarte e Serra, 2015: 42).

Como salientam Duarte e Serra (2015), o problema do uso exclusivo de gramáticas tradicionais no ensino é que muitas delas "têm sido reeditadas sem incorporar os avanços dos estudos linguísticos, sem sanar a inconsistência/incoerência dos conceitos utilizados na descrição da língua e sem atualizar os dados relativos ao uso normal da escrita" (Duarte e Serra, 2015: 37). Em vista disso, muitas das regras apresentadas pelas gramáticas tradicionais não fazem parte da gramática genuinamente brasileira, mas podem estar presentes na língua escrita formal.

Diante disso, há um enorme fosso entre a escrita dos brasileiros escolarizados, dita "norma culta", e a fala popular daqueles que não tiveram acesso ao ensino. Segundo Lucchesi, a língua popular brasileira passou a sofrer um estigma social "e o *preconceito linguístico* constitui um poderoso mecanismo ideológico de legitimação da exclusão social e da exploração do trabalho" (2012: 50). A língua falada por brasileiros não escolarizados é altamente estigmatizada, mesmo com o surgimento de diversas pesquisas linguísticas que comprovam as semelhanças entre a variação presente na língua falada do português culto e do português popular (Duarte, 2015: 23-24).

Essa discriminação de formas linguísticas genuinamente brasileiras leva à ideia equivocada de que a distância entre a norma culta e a fala popular seria uma questão de "formalidade *versus* informalidade", em que o PE seria considerado mais formal e o PB mais informal (Duarte, 2015: 28). Acerca disso, Duarte e Serra (2015:43) ressaltam a importância de não confundir "adequação de linguagem" com mudança de gramática em situações formais. Quando um falante que domina a norma culta deixa de usar a variante popular da língua para empregar estruturas mais próximas do PE em situações formais, ele não utiliza simplesmente um mecanismo de adequação de linguagem: nesses contextos, brasileiros escolarizados efetuam uma mudança de gramática, optando por estruturas que foram objeto de aprendizagem explícita do português em contexto escolar.

Para iniciar a discussão acerca da descrição do PB padrão é necessário evidenciar o que se entende por norma culta e escrita padrão. Por um lado, a norma culta é a variante perpetuada pela gramática normativa, segundo a qual "a língua corresponde às formas

de expressão produzidas por pessoas cultas, de prestígio" (Possenti, 1996: 54). Por outro lado, a escrita padrão é definida como "as variedades de escrita veiculada em jornais e revistas de ampla circulação, em trabalhos acadêmicos, em escrita produzida por indivíduos escolarizados e com contato frequente com a escrita" (Duarte e Serra, 2015: 39). Apesar de a escrita padrão corresponder, na sua maioria, à norma culta, devido à posição de prestígio dessa variante e pelo fato de a escrita ser utilizada por falantes escolarizados que têm acesso a ela, a língua padrão apresenta maior variação do que a língua prescrita pelas gramáticas normativas, pois corresponde ao que é efetivamente escrito pelos brasileiros escolarizados.

Segundo Perini (1985), a elaboração da gramática do PB deve ser feita a partir da descrição de estruturas presentes em textos escritos em linguagem não-literária. A partir dessa ideia, Duarte (2015) explicita que essa descrição deve ser plural, dado o contexto de variação da escrita padrão do PB que possui características próprias, mas ao mesmo tempo manifesta traços do PE, difundidos pela tradição gramatical. De acordo com essa autora,

> "[é] imperioso dar o primeiro passo: reunir os resultados de que já dispomos e ampliá-los com base em dados efetivos da escrita contemporânea, sem excluir ou condenar esta ou aquela forma (fazer isso seria incorrer no mesmo erro do passado), sem negar que *essa escrita é variável*, que essa variação está nos nossos artigos acadêmicos, nas teses e dissertações que orientamos, nos textos dos meios de comunicação escrita mais prestigiados do país" (Duarte, 2015: 36).

O estudo de Kato (2005) acerca da gramática do letrado também confirma a ideia de que, apesar dos esforços do ensino da gramática normativa em promover uma língua homogênea com normas próximas a gramática do PE, ao seu ritmo, a língua escrita na variedade brasileira reflete a variação já presente na língua falada, apresentando simultaneamente estruturas do modelo lusitano e do PB contemporâneo.

A existência de uma variação entre a língua oral e escrita em PB é inegável e já foi comprovada em diversos estudos feitos nos últimos anos. A presente pesquisa, tomando as ideias de Duarte (2015) acerca da necessidade de uma descrição plural do PB, procura criar um diálogo entre as regras apresentadas nas gramáticas normativas e o uso efetivo de estruturas do PB em textos escritos em português padrão, a fim de analisar os erros de concordância e ordem de palavras presentes nos dados da Unbabel.

No intuito de respeitar a ideia de uma descrição plural e alcançar os objetivos desta pesquisa, considerou-se as definições e ideias de pesquisas linguísticas presentes na

literatura sobre o assunto em PB, bem como a literatura em PE, que possa auxiliar na elaboração dessa descrição. Para mais, por motivos de economia, somente foram apresentados nas descrições os fenômenos de concordância e ordem de palavras relacionados com os erros encontrados nos dados em PB fornecidos pela Unbabel. Por essas razões, as próximas seções não consideram as descrições que tratam da língua oral em PB, tendo em vista que os dados a analisar no âmbito desta pesquisa são todos dados escritos. A maior parte dos exemplos apresentados nas subseções a seguir são excertos de textos jornalísticos da *Folha de São Paulo*, escritos em PB, encontrados através da base de dados do CETENFolha, um dos *corpora* acessíveis através do Projeto AC/DC: *Corpus* NILC/São Carlos, no site da Linguateca.

5.1 Questões sobre a concordância em PB

A presente seção se dedica à descrição geral dos fenômenos de concordância em PB, observando os casos em que se encontram envolvidas classes variáveis[9] interessantes para a análise dos dados. Antes de iniciar essa discussão, é oportuno introduzir brevemente o conceito de predicador e a sua importância nas relações de concordância no nível frásico. Segundo Raposo (2013: 358), "o predicador de uma frase é o item lexical que define o conteúdo fundamental das proposições, independentemente da sua natureza semântica mais precisa como representando ações, atividades, processos ou situações estáticas".

Ainda segundo esse autor, podem ser predicadores os verbos (1a), os adjetivos (1b), os nomes (1c), determinadas preposições (1d) e certos advérbios (1e). Com exceção dos verbos, os termos das outras classes gramaticais funcionam como predicadores em orações com verbo copulativo, em que os verbos "embora veiculem valores semânticos importantes nas frases, não funcionam como predicadores, visto não terem um conteúdo descritivo" (Raposo, 2013: 359).

(1)

a. A noiva *lavou* o vestido. (verbo)
b. A Joana é *teimosa*. (adjetivo)
c. A Luísa é (uma) *enfermeira*. (nome)
d. O livro está *sobre* a mesa. (preposição)
e. O espetáculo foi *ontem*. (advérbio)

(Raposo, 2013: 359)

[9] Considerando-se a divisão feita por Raposo (2013: 332) entre as classes de palavras variáveis (verbos, nomes, adjetivo, determinantes, quantificadores, pronomes e certos numerais); e invariáveis (preposição, advérbio, conjunção e interjeição).

Como é possível verificar nas frases acima, o predicador tem o papel de núcleo semântico da frase e do predicado: é ele o elemento semântico mais importante da frase (Raposo, 2013: 359). Quando os núcleos predicativos pertencem a classes variáveis, ou seja, adjetivos, verbos ou nomes, os predicadores devem compartilhar os traços de pessoa (no caso do verbo), gênero (no caso do adjetivo e, geralmente, do nome) e número (no caso do verbo, do adjetivo e do nome) com o sujeito da frase, como pode ser observado nos exemplos (1a), (1b) e (1c). As preposições e advérbios, quando são predicadores, não participam da concordância, como em (1d) e (1e).

Considerando essas definições, a presente seção foi dividida em três seções. Na primeira (5.1.1), é abordada a concordância no interior do SN**,** logo, será observada a flexão de nomes, adjetivos, certos pronomes, quantificadores e determinantes. Na segunda (5.1.2), são explicitadas as relações de concordância envolvendo sujeitos, detalhando a relação de concordância entre o sujeito e o verbo (finito e não finito) e entre o sujeito e o predicativo (em orações copulativas); também nessa seção será mencionado o funcionamento do pronome relativo quanto a esse aspecto. Na terceira (5.1.3) será tratada a ligação entre a concordância e a coesão textual, explicitando a importância do compartilhamento de traços de gênero, número e pessoa no estabelecimento de relações anafóricas e dêiticas.

5.1.1 Concordância no interior do SN

O sintagma nominal (doravante SN) é sempre constituído por um núcleo e pode estar acompanhado de especificador, modificador(es) e complemento(s), como exemplificados em (2). As relações de concordância entre esses elementos e o núcleo do SN dependem das classes gramaticais que os constituem. Por exemplo, em (2a) o especificador "*os*" e o modificador "*doces*" concordam em gênero e número com o núcleo do SN no qual estão inseridos. Já no caso do complemento preposicionado em (2b) não há concordância entre o núcleo "diminuição" e o complemento "de doações", inseridos no SN.

(2)

a. [especificador – núcleo – modificador]: **Os alimentos doces**...
b. [especificador – núcleo – complemento]: ...**a diminuição de doações**...

(Corpus: CETENFolha)

Em PB, a flexão de plural na língua oral é frequentemente feita somente nos elementos mais à esquerda do SN (cf. Duarte, 2015; Duarte e Serra, 2015; Scherre e Naro 1998b).

Porém, na escrita ainda há uma enorme preferência pela marcação redundante da concordância, não sendo possível encontrar estruturas como "*as menina bonita*" em textos jornalísticos, exceto em contextos específicos em que se procura aproximar o texto escrito da língua popular ou em gêneros textuais mais literários, como as crônicas jornalísticas. Os parágrafos seguintes procuram explicitar as relações de concordância no interior do SN relevantes para a análise dos dados em PB escrito.

5.1.1.1 Concordância entre nome e especificadores

No que diz respeito ao artigo definido, esse especificador determinante do nome varia em gênero e número, concordando com o núcleo do SN. Ele se flexiona nas quatro formas *o*, *a*, *os*, *as*, (3a)[10] que podem se contrair com as preposições *de*, *a*, *em* e *por*, formando combinações como *do*, *ao*, *no* e *pelo* (e suas variantes). Essas combinações podem ser observadas nos seguintes exemplos: (3a) ***ao*** *Planejamento*; (3b) ***do*** *time*... ***pela*** *classificação*...***à*** *Copa*; e (3c) ***nos*** *telejornais*, ***nas*** *eleições*. O excerto (3c) exemplifica o uso desse determinante no plural, combinado (cf. *nos*, *nas*) ou não (cf. *as*) com preposição.

(3)

a. **O** tribunal, diz, teria mandado a tempo **a** proposta **ao** Planejamento.
b. Cada jogador **do** time ganhou US$ 100 mil **pela** classificação **à** Copa.
c. **As** manchetes, **nos** telejornais, só falavam em tranquilidade, em festa ou em calma, **nas** eleições.

(*Corpus*: CETENFolha)[11]

O artigo indefinido também varia em gênero e número, concordando com o núcleo do SN. Ele apresenta as seguintes formas: *um*, *uma*, *uns* e *umas*. Os elementos assinalados em (4a), "*uns*", "*umas*" e "*um*" exemplificam esse tipo de artigo. Quando combinadas com as preposições *em* e *de*, essas formas se contraem: *num*, *dum*, *nuns* e *duns* (e variantes), como o exemplo "*num*" em (4b).

(4)

a. ...levo **uns** limões galegos para fazer caipirinha, **umas** caixas de fósforos para garantir a batucada e **um** berimbau...
b. ...está **num** documento entregue...

(Corpus: CETENFolha)

[10] Por razões de economia, os exemplos apresentados nesta descrição incluem apenas algumas das formas.
[11] Foi inserida nesta pesquisa somente o trecho considerado relevante para ilustrar a descrição. Para mais contexto, cf. Corpus NILC/São Carlos: CETENFolha, disponível em https://www.linguateca.pt

O português possui ainda três grupos de determinantes demonstrativos: *este*, *esse* e *aquele.* Cada um desses grupos inclui suas variantes em gênero em número. Em (5a), é possível verificar que o demonstrativo "*esse*" concorda com o seu núcleo masculino singular "*estágio*". Em (5b) o demonstrativo "*esta*" concorda com a palavra feminina singular "*safra*". As formas neutras de cada grupo de determinante demonstrativo (*isto*, *isso*, *aquilo*) não são utilizadas como determinantes, mas sim como pronomes.

(5)

a. ...para **esse** estágio.
b. Para **esta** safra....

(Corpus: CETENFolha)

Quando os SN iniciados por esses demonstrativos são introduzidos pelas preposições *de* e *em*, eles se contraem formando *deste*, *desse*, *daquele* (e variantes); e *neste*, *nesse*, *naquele* (e variantes). A forma "*desse*" em (6a) exemplifica a combinação da preposição *de* e do demonstrativo masculino singular *esse*. Em (6b), a forma demonstrativa feminina singular *esta* se combina com a preposição *em*, formando a combinação "*nesta*". No caso da contração com a preposição *a*, somente o grupo *aquele* se contrai: *àquele* (e variantes), como no exemplo (6c).

(6)

a. Parte **desse** dinheiro...
b. ...enquadra-se **nesta** definição.
c. ...o repúdio **àquele** tributo.

(Corpus: CETENFolha)

Acerca do uso dos possessivos como especificadores do SN, Raposo e Miguel (2013: 730) ressaltam uma diferença entre PE e em PB: em PB, assim como em inglês (*my book*), francês (*mon livre*) e outras línguas românicas, o possessivo ocorre na posição inicial do SN geralmente sem determinante (*meu livro*). Considerando-se que esse SN possui uma interpretação definida nessas línguas, "é legítimo considerar que os pronomes possessivos funcionam como determinantes (sem deixarem de funcionar como complementos ou modificadores)" (Raposo e Miguel, 2013: 730). Castilho (2010: 530) também insere os possessivos na classe dos especificadores do SN. Não é o objetivo da presente pesquisa entrar em detalhes acerca dessa classificação, mas acolhemos essas

afirmações ao longo da presente descrição. Os excertos (7a) e (7b) apresentados a seguir exemplificam a possibilidade do uso de possessivos pré-nominais não acompanhados de outro especificador:

(7)

a. **Nossa** economia (...) a qualidade de **nossas** demandas...
b. ...sugeria **minha** eliminação...

(Corpus: CETENFolha)

Quanto à sua morfologia, os possessivos se flexionam em pessoa, número e gênero. No traço de pessoa, o possessivo concorda com a entidade "possuidora", já nos traços de número e gênero ele deve concordar com o núcleo do SN (entidade possuída). As formas do possessivo apresentam uma forma feminina (singular e plural) e masculina (singular e plural) para cada pessoa gramatical. Em (8a), por exemplo, o pronome possessivo "*minhas*" concorda com o substantivo feminino plural "*frustrações*", núcleo do SN, e está na 1ª pessoa do singular. O mesmo ocorre em (8b), em que "*vosso*" concorda com "*tio*", mas o possuidor não é o próprio falante.

(8)

a. ...falava das **minhas** frustrações...
b. Quando **vosso** tio Dave se casou…

(Corpus: CETENFolha)

Além desses determinantes, o SN pode ser especificado por elementos que expressam uma ideia de quantidade. Fazem parte desse conjunto os quantificadores universais, como *todos* (9a) e *nenhum* (9b) e suas variantes femininas, e os quantificadores "vagos", como *bastante* (10a) e *poucos* (10b) e suas variantes femininas (Raposo e Miguel, 2013: 718). Os primeiros se referem à totalidade das entidades de um determinado conjunto, enquanto os segundos possuem um valor impreciso. Os quantificadores devem concordar em gênero e número com o núcleo do SN . Por isso, no português escrito não seria possível construções como **todos equipamento* em (9a) ou **bastantes espaço* em (10a).

(9)

a. ...**todos** equipamentos...
b. ...em **nenhuma** localidade.

(10)

a. ...dedique **bastante** espaço...

b. ...ofertou **poucos** animais puros...

(Corpus: CETENFolha)

A classe dos numerais é bastante heterogênea, uma vez que inclui diversos elementos que denotam a ideia de número, mas que fazem parte de classes de palavras distintas (Vicente, 2013: 925). Interessa-nos os numerais cardinais comuns podem ser especificadores quantificacionais do SN, apresentando propriedades semelhantes às dos quantificadores "vagos" e universais. Esse tipo de numeral é invariável em gênero, com exceção de *um*/*uma* e *dois*/*duas*, mas expressa uma quantidade (singular ou plural) que deve concordar com a do seu núcleo, por isso em (11a) o traço de número em "*cinco*" concorda com o plural de "*parcelas*" e em (11b) a quantidade expressa no numeral "*vinte*" concorda com a quantidade plural de "*fascículos*". Além disso, em (11a) uma construção como "***cinco*** *parcela mensal*", apesar de possível no PB oral, não seria adequada na variedade escrita.

(11)

a. ...em **cinco** parcelas mensais...
b. ...com **vinte** fascículos...

(Corpus: CETENFolha)

5.1.1.2 Concordância entre nome e modificadores

Segundo Brito e Raposo (2013: 1045), os modificadores do nome se inserem no SN para introduzir propriedades adicionais na denotação de um nome, ou de um nome com seu complemento. Eles restringem o sentindo do nome, dando-lhe maior precisão. Podem ser modificadores do nome os sintagmas adjetivais (doravante SAs) com núcleo adjetivo (12a) ou adjetivos derivado de particípio (12b); as orações relativas (12c); os possessivos (12d); e os SPs (12e).

(12)

a. É uma farmácia **verde**, **imensa** e **fascinante**.
b. ...os candidatos **derrotados**...
c. A cultura **que flutuava num progresso feliz**...
d. ...um sonho **meu**.
e. ...no aeroporto **de Argel**...

(Corpus: CETENFolha)

Como é possível verificar a partir dos exemplos apresentados, somente não concordam com o núcleo nos traços de gênero e número o SP "*de Argel*" em (12e). Os modificadores oracionais constituem um caso particular, pois o seu comportamento se insere na concordância verbal: o verbo se flexiona em número e pessoa de acordo com o

seu argumento externo, que corresponde tipicamente a um pronome relativo que recupera o SN (Brito e Raposo, 2013: 1103). Tendo vista essas propriedades, os modificadores oracionais serão tratados na seção 5.1.2 referente à concordância verbal.

Como é possível verificar nos exemplos seguintes, independentemente de o modificador adjetival ocorrer antes (13a) ou depois do núcleo nominal (13b), a concordância em gênero e número é obrigatória. Mais detalhes acerca das circunstâncias nas quais os adjetivos ocorrem nessas posições serão fornecidos na seção 5.2.2.2, referente à ordem dos modificadores no interior do SN.

(13)

a. ...como uma **barulhenta** chuva de verão?
b. ...para a audiência **barulhenta**.

(Corpus: CETENFolha)

Quanto ao particípio modificador do SN, é de assinalar que, "tal como os adjetivos, a variação em gênero e número dos particípios resulta da concordância (...) com o nome que modificam atributivamente" (Veloso e Raposo, 2013: 1477). Os particípios assinalados em (14a) e (14b) concordam em gênero e número com os núcleos do SN, "*trabalhadores*" e "*títulos*", respectivamente.

(14)

a. ...todos os trabalhadores **registrados** do país.
b. ...dos títulos **emitidos**.

(Corpus: CETENFolha)

Como já foi visto, segundo Raposo e Miguel (2013: 730) os possessivos pré-nominais exercem a função de especificadores do SN em PB. No caso dos possessivos pós-nominais, Raposo (2013: 906) os classifica de forma distinta segundo as propriedades do núcleo do SN, funcionando como **complementos** quando o núcleo é um nome dependente, ou seja, que pode selecionar um complemento (15a); e como **modificadores** quando o núcleo é um nome autônomo (15b).

(15)

a. Os companheiros **nossos** têm...
b. ...um livro **seu**, assim, dá...

(Corpus: CETENFolha)

Não entra nos objetivos deste trabalho a extensa diferenciação entre esses três tipos de possessivos. Em todos esses casos, os possessivos devem seguir as mesmas regras de concordância já citadas na seção referente aos especificadores possessivos. Mais detalhes acerca da diferença entre complementos e modificadores do SN serão fornecidos nos parágrafos seguintes.

5.1.1.3 Concordância entre nome e argumentos adjetivais

Para discutir a natureza dos complementos, é necessário primeiramente fazer uma distinção entre nome autônomo e dependente: os nomes do primeiro grupo denotam entidades imediatamente identificáveis sem a necessidade de relacioná-las com outras entidades, por isso não selecionam complementos, como *mesa*, *cavalo*, *pão*, *madeira*; os do segundo grupo denotam tipicamente entidades do mundo que só podem ser apreendidas quando são postas em relação com outras entidades, como *casamento* (de alguém), *autor* (de uma obra) e *consequência* (de algo), por isso selecionam complementos (Raposo e Miguel, 2013: 715). Os nomes assinalados nos exemplos a seguir exemplificam um nome autônomo (16a) e um nome dependente (16b), respectivamente: o termo "*cirúrgica*" é um modificador e a expressão "*de homossexuais*" é um complemento.

(16)

a. ...para a **mesa** *cirúrgica*.
b. Legalização do **casamento** *de homossexuais*.

(Corpus: CETENFolha)

Esses autores ressaltam que os nomes deverbais, relacionados com verbos ou derivados deles, possuem um comportamento particular, visto que herdam os argumentos dos verbos dos quais derivam. Em (17a), a palavra "*destruição*", derivada do verbo *destruir*, seleciona o complemento "*de Pompéia*". Em (17b), dois complementos são selecionados pelo nome "*entrada*": "*de mangas*" e "*no país*".

(17)

a. A **destruição** [*de Pompéia*] foi tão brusca...
b. Para autorizar a **entrada** [*de mangas*] [*no país*]...

(Corpus: CETENFolha)

Isto posto, os complementos são expressões de natureza argumental que completam o sentido dos nomes (Raposo e Miguel, 2013: 715). Apesar das diferenças entre esses argumentos adjetivais e os modificadores, as relações de concordância entre ambos e o núcleo do SN que integram não são muito diferentes: quando são SA, (18a), concordam com o núcleo do SN em gênero e número.

(18)

a. ...a possibilidade de uma invasão **política**.

(Corpus: CETENFolha)

Os argumentos do nome que não são adjetivais, em particular, SPs introdutores de orações completivas (19) ou de SNs, como *uma invasão política* em (18), não estabelecem qualquer relação de concordância com o nome.

(19)

a. ...descartou a *possibilidade* **de que *seja decidido* na reunião um novo ataque a Base.**
b. ...afastam a *possibilidade* **de *seguir* o mesmo rumo de Vasconcelos.**

(Corpus: CETENFolha)

Como resultado, é possível concluir que o SN apresenta diversas possibilidades de estruturação, podendo o seu núcleo ocorrer isolado ou acompanhado por especificadores, modificadores e complementos. Além disso, foi possível verificar que as relações de concordância estabelecidas entre o núcleo e esses elementos no interior do SN dependem das características desses elementos.

5.1.2 Concordância envolvendo Sujeito

Como já foi mencionado no início da presente seção, podem ser predicadores o verbo, o adjetivo, certas preposições e alguns advérbios. Os dois últimos tipos de predicadores não serão abordados nesta seção, pois são classes invariáveis. Serão então apresentadas as relações de concordância entre sujeitos e predicadores verbais, nominais e adjetivais. De maneira geral, no caso da predicação verbal, o sujeito concorda em pessoa e número com o verbo, que é tipicamente o predicador da frase , e no caso da predicação nominal e adjetival, o sujeito concorda com o verbo copulativo e com o predicativo (composto por SA ou SN).

Para explicitar ainda mais essas relações de concordância, apresentamos os seguintes exemplos retirados do *corpus* consultado: em (20a) o sujeito (*os italianos*) partilha os traços de pessoa e número com o verbo predicador da oração (*recuperaram*); em (20b) todo o SN (*pessoas autoconfiantes*), predicativo do sujeito, concorda em gênero e em número com o sujeito (*empreendedores*); em (20c) o SA (*maravilhosos*), predicativo do sujeito, concorda em gênero e número com o sujeito masculino plural (*os enterros no Brasil*). Também os verbos copulativos presentes nas orações (20b) e (20c) concordam com os sujeitos da frase.

(20)

a. Os italianos **recuperaram** o controle...
b. Empreendedores são **pessoas** autoconfiantes...
c. Os enterros no Brasil são **maravilhosos**.

(Corpus: CETENFolha)

A presente seção discute os fenômenos de concordância entre o sujeito e o verbo na seção 5.1.2.1, subdividida em estruturas com verbo finito (5.1.2.1.1) e com verbo não finito (5.1.2.1.2); e entre o sujeito e o predicativo na seção 5.1.2.2. Tendo em conta os objetivos da presente pesquisa, na subseção de verbos finitos serão mencionadas a variação da concordância verbal em PB e o fenômeno da concordância semântica, já na subseção de verbos não finitos foi dada atenção especial aos casos de infinitivo flexionado.

5.1.2.1 Sujeito – Verbo

Segundo Raposo (2013: 329), os verbos possuem uma forma finita e uma não finita. Na sua forma finita, essa classe gramatical possui desinências que expressam as informações semânticas de tempo (21a), modo (21b), aspecto (21c), pessoa (21d) e número (21e) como é possível verificar nas diferentes desinências do verbo *cantar*, apresentadas a seguir.

(21)

a. Tempo: cant*o* – cant*ei* – cant*arei*
b. Modo: cant*o* – cant*e*
c. Aspecto: cant*ei* – cant*ava*
d. Pessoa: eu cant*o* – tu cant*as*
e. Número: eu cant*o* – nós cant*amos*

(Raposo, 2013: 329; com o acréscimo de (21e))

A concordância verbal é o fenômeno em que as marcas flexionais de pessoa e número do verbo são controladas pelo sujeito. Logo, participam da concordância entre o sujeito e o verbo os verbos finitos e os verbos no infinitivo flexionado. Contudo, é de notar a diferença entre os verbos pessoais e os verbos impessoais: enquanto os verbos do primeiro grupo selecionam um sujeito gramatical com o qual partilham traços de pessoa e número, os verbos impessoais não selecionam um argumento com a função de sujeito gramatical (ou seja, a posição de sujeito é nula) e "conjugam-se invariavelmente na 3ª pessoa do singular" (Gonçalves e Raposo, 2013: 1193-1195).

Nos exemplos seguintes, os verbos impessoais *choveu* e *havia* estão ambos na 3ª pessoa do singular em (22). Já em (23), os verbos pessoais *prometeu* e *cultiva* foram conjugados na 3ª pessoa do singular segundo os traços do sujeito expresso (*ele* em (23a)) ou do sujeito nulo (representado por "[-]" em (23b)).

(22)

a. **Choveu** bastante sábado na cidade.
b. **Havia** algo de Gandhi no Betinho.

(23)

a. *Ele* **prometeu** também ajuda militar...
b. [-] **Cultiva** o sentimento da ambiguidade...

(Corpus: CETENFolha)

Antes de apresentar mais detalhes acerca das estruturas de concordância verbal em PB, é necessário apontar a existência da concordância por sentido (ou concordância semântica) em português, já reconhecida pela tradição gramatical. É possível encontrar em Cunha e Cintra (1985: 350-360), dentro de uma extensa lista de casos particulares, a possibilidade da concordância por sentido em português. Também Bechara aponta que "a concordância pode ser estabelecida de palavra para palavra ou de palavra para sentido. (...) A concordância de palavra para sentido se diz ainda concordância '*ad sensum*' ou silepse" (1961/2002: 543).

Ao observarem textos escritos no português padrão, Scherre e Naro (1998a) assinalam que "diferentemente da língua falada, na língua escrita moderna, a concordância verbal de número plural é quase categórica com sujeito simples (de um só núcleo) de estrutura simples (sem sintagmas preposicionais – SPs – adjunto ou complemento)" (*op. cit.*: 49). Os sujeitos simples das orações apresentadas a seguir exemplificam esse tipo de estrutura com concordância singular (24a) e plural (24b):

(24)

a. *A obra* **anunciava** o estilo e o conteúdo...
b. *As máquinas* **imprimem**, também, logotipos...

(Corpus: CETENFolha)

Ainda segundo esses autores, a variação ocorre de maneira regular na escrita moderna "em estruturas de sujeito simples de estrutura complexa, cuja configuração sintagmática se apresenta na forma de um núcleo, seguido de sintagma preposicional" (Scherre e Naro, 1998a: 49). Eles apontam as construções em que se registra regularmente variação possuem um sujeito de núcleo singular de natureza quantitativa seguido de um SP de núcleo plural, fazendo assim com que o sujeito denote "uma leitura quantitativa, coletiva ou partitiva" (Scherre e Naro, 1998a: 49).

É possível observar essa variação nos seguintes excertos de textos jornalísticos escritos em PB, retirados do texto dos autores, em que todos os sujeitos possuem um SP plural e um núcleo singular de natureza quantitativa, tendo em vista a presença de *grupo* em (25) e *maioria* em (26), mas o verbo se flexiona no singular em (25a) e (26a) e no plural em (25b) e (26b).

(25)

a. Um *grupo* de artistas **estava** sábado...
b. Um *grupo* de "homens da cidade" **decidem** ir atrás do ouro...

(26)

a. A *maioria* dos deputados **encenou** um espetáculo...
b. A *maioria* dos pais **impõem** restrições...

(Scherre e Naro, 1998a: 49)

Segundo Raposo (2013: 975), também em PE "os nomes coletivos no singular, quando funcionam como sujeito, determinam frequentemente concordância verbal no plural". Além da ocorrência de concordância plural com SN coletivo em que se encontra um SP plural, já apresentadas nos exemplos anteriores, também é possível encontrar ocorrências de plural quando o SN contendo nome coletivo possui um SP reduzido singular (27a) ou não possui um SP reduzido (27b), ocorrendo nesse caso a concordância com o sentido semântico plural do sujeito, como nos seguintes exemplos retirados do texto do autor:

(27)

a. Uma *brigada* da PSP, trajando à civil, **detiveram** anteontem...
b. É um *grupo pop* e **cantavam** cânticos negros.

(Raposo, 2013: 976)

Scherre e Naro (1998a) prosseguem na investigação acerca da concordância verbal em PB, defendendo a hipótese de que o SP do sujeito complexo pode assumir o controle da concordância[12], pois se registram casos na língua escrita moderna "em que se observa a concordância formalmente plural ou formalmente singular com o núcleo do sintagma preposicional" (Scherre e Naro, 1998a: 50), independentemente de o núcleo do sujeito não ter sentido quantitativo (28a) ou ser plural (28b), como é possível verificar nos seguintes exemplos retirados dos textos dos autores:

(28)

a. A *construção* de mais três escolas **estão** nos planos...
b. As *mudanças* bruscas do momento político **pode** provocar...

(Scherre e Naro, 1998a: 50)

Scherre e Naro (1998a: 50-56) sugerem que o traço semântico [+ ou – humano] influencia na variação da concordância verbal. Essa influência é confirmada na pesquisa sobre a concordância em PE e PB de Monguilhott e Coelho (2011: 310), na qual se verificou uma frequência maior de concordância quando o sujeito tem traço [+ humano], como em (29a), por oposição a sujeitos com traços [- humano], como em (29b). Tendo em vista a extensão e os objetivos da presente pesquisa, apesar de se reconhecer a existência dessa influência, optou-se por não detalhar ainda mais essa hipótese.

(29)

a. Todas *as minhas amigas* **namoravam** e **vinham** às festas...
b. Tem *várias etnias* que **contribuíram** pra formação...

(Monguilhott e Coelho, 2011: 310)

Diversas pesquisas acerca da variação no PB demonstram que a posição do sujeito em relação ao verbo também influencia na variação da concordância verbal: "uma leitura mais cuidadosa evidenciará a não-realização da concordância por conta da posposição do elemento que a controla – o sujeito, e por vezes, da distância entre sujeito e verbo" (Faria, 2017: 155). O mesmo é observado do por Monguilhott e Coelho (2011) ao analisarem a frequência da marcação de concordância verbal em textos escritos em PB e PE. Também

[12] Esses autores apontam a influência do SP na variação da concordância verbal também em construções que expressam percentual (Scherre e Naro, 1998a: 52). Esses casos não serão mencionados, considerando-se a extensão e os objetivos da presente dissertação.

Castro (2016), ao analisar a concordância verbal na escrita universitária em PB, observa o favorecimento da concordância com sujeitos antepostos. Os seguintes excertos retirados do texto desses autores exemplificam o uso da concordância verbal quando o sujeito é anteposto ao verbo (30a) e a falta de marcas de concordância de número quando o sujeito é posposto (30b).

(30)

a. *As mulheres* não **tinham** direito a voto

(Monguilhott e Coelho, 2011: 310)

b. ...já completando um ano que **surgiu** *as primeiras informações* que a prefeita poderia ser afastada do cargo...

(Faria, 2017: 155)

A partir do exposto, é possível concluir que o não compartilhamento de traços entre o núcleo do sujeito e o verbo, considerado como uma manifestação rara e esporádica no português, na verdade se apresenta de forma recorrente segundo a estrutura do sujeito e a posição desse elemento na oração. Na prática, quando se observam as variadas possibilidades de concordância, é possível concluir que há uma forte conexão entre a variação e a noção de concordância semântica.

Para finalizar, é possível afirmar que há três diferentes tipos de concordância em português: "(1) concordância gramatical: um termo concorda com o outro; (2) concordância semântica: um termo concorda com o sentido de outro; (3) concordância discursiva: um termo do enunciado concorda com um participante da enunciação" (Castilho, 2010: 272-273). Esses três tipos de concordância são ilustrados a seguir, respectivamente, em exemplos do mesmo autor. No caso de (31a), o verbo concorda gramaticalmente com o singular expresso por *a totalidade*, em (31b) o verbo concorda com o sentido plural de *multidão* e em (31c) a palavra *encantada* concorda com o gênero feminino do falante.

(31)

a. *A totalidade dos entrevistados* **deixou** de aparecer.
b. *A multidão*, depois do cerco da polícia, **foram** saindo de fininho.
c. *Eu* fiquei **encantada** com tudo aquilo.

(Castilho, 2010: 272-273)

Considerando a existência desses três tipos de concordância, será dada atenção especial à concordância discursiva na seção 5.1.3 acerca da concordância e coesão textual. A concordância semântica será lembrada durante a análise dos dados, ainda assim, para cumprir com os objetivos desta pesquisa, optou-se por apresentar nos

parágrafos seguintes uma descrição geral das estruturas de concordância gramatical em PB escrito relacionadas com o que é relevante nos dados a serem analisados, deixando de lado, por agora, outras hipóteses de variação.

5.1.2.1.1 Verbos finitos

Na predicação verbal, o verbo (principal) é o elemento semântico central da oração. Logo, quando desacompanhado de verbo auxiliar, é o verbo pleno que partilha as marcas de pessoa e número com o sujeito da oração (Gonçalves e Raposo, 2013: 1155). A flexão de verbos plenos nas 3 pessoas gramaticais no singular e plural pode ser observada nos seguintes exemplos retirados do *corpus* consultado.

(32)

a. 1ªpsg: Eu **diria** que 75 %...
b. 2ªpsg: ...tu **pedes** um milhão...
c. 3ªpsg: ...ele **visitou** as cidades...
d. 1ªppl: ...nós **sabemos** disto...
e. 2ªppl: Vós **sumistes**!
f. 3ªppl: ...elas **estipularam** taxas...

(Corpus: CETENFolha)

É importante ressaltar que o paradigma pronominal está em processo de mudança em PB (cf. Mattos e Silva, 2013). Assim, a utilização dos sujeitos nominais *você*, *vocês* e *a gente*, ao invés dos pronomes *tu*, *vós* e *nós* é cada vez mais frequente. Com essas formas nominais, semanticamente o sujeito tem sentido de 2ªpsg (*você* em (33a)), 2ªppl (vocês em (33b)) ou 1ªppl (*a gente* em (33c)), mas gramaticalmente o verbo se conjuga na 3ª pessoa do singular ou plural, como é possível observar nos exemplos retirados do *corpus* consultado.

(33)

a. 3ªpsg: Se *você* **come** carne...
b. 3ªppl: Agora, *vocês* **ganham** lá...
c. 3ªpsg: *A gente* **precisa** de reciclagem...

(Corpus: CETENFolha)

Os verbos plenos podem ser acompanhados por verbos auxiliares ou semiauxiliares, constituindo-se assim uma perífrase verbal. Segundo Raposo, ao se combinarem com os verbos plenos, esses verbos "contribuem com informação nos domínios semânticos do tempo, da modalidade e do aspecto" (2013: 1221). Apesar dessa contribuição, os verbos plenos continuam sendo o elemento semântico central da oração e verbos auxiliares não

podem ser predicadores. Isso pode ser observado nos seguintes exemplos, retirados do texto desse autor: os verbos plenos *chover* (34a), *vir* (34b) e *fazer* (34c) são precedidos de verbos auxiliares sobre os quais recaem os traços de tempo, modo e aspecto, bem como a concordância de pessoa e número. Todavia, o conteúdo semântico crucial desses excertos está presente nos verbos plenos.

(34)

a. **Vai chover**. (domínio do tempo)
b. A Clara **pode vir** a Lisboa hoje. (domínio da modalidade)
c. O meu carro **continua a fazer** um barulho esquisito. (domínio do aspecto)

(Raposo, 2013: 1221)

Esse autor também aponta que uma perífrase verbal contém somente um verbo pleno, mas pode incluir mais do que um verbo auxiliar (Raposo, 2013: 1225). Nesse caso, o verbo pleno ocorre na última posição da perífrase verbal e, nos casos em que a frase é finita, somente o primeiro verbo da sequência se flexiona em tempo, modo, aspecto e concorda com o sujeito em número e pessoa (Raposo, 2013: 1226). Com exceção do primeiro verbo auxiliar, "todos os outros verbos, incluindo o verbo pleno, ocorrem numa forma não finita, que pode ser o infinitivo, o particípio ou o gerúndio" (Raposo, 2013: 1226).

Essas afirmações de Raposo (2013) podem ser confirmadas nos exemplos retirados do *corpus* consultado: na última posição da perífrase verbal estão sublinhados os verbos plenos no infinitivo (35a), no particípio (35b) e no gerúndio (35c), detalhes acerca do funcionamento dessas formas não finitas serão dados na próxima seção (5.1.2.1.2). Também é possível verificar que os traços de concordância verbal com o sujeito da oração recaem sobre o primeiro verbo da sequência: *vamos* (35a), *pode* (35b) e *teriam* (35c).

(35)

a. ...não ***vamos*** *conseguir* <u>*competir*</u> com a eficiência. (infinitivo)
b. ...a droga ***pode*** *ter* <u>*bloqueado*</u> a irrigação de sangue... (particípio)
c. ...sequestradores ***teriam*** *escapado* <u>*levando*</u> Silva. (gerúndio)

(Corpus: CETENFolha)

Também entram nas relações de concordância verbal as orações relativas, que tipicamente "modificam um nome ou grupo nominal dentro de um SN complexo" (Veloso, 2013: 2061). Como já foi referido na seção 5.1.1, nesse caso a concordância está relacionada com as regras de concordância verbal: o verbo se flexiona em número e pessoa de acordo com o seu argumento externo (Brito e Raposo, 2013: 1103). Logo,

mesmo com sujeito não expresso, os verbos *vês* (36a) e *comprei* (36b) não partilham os mesmos traços de pessoa e número dos nomes *rapariga* e *livros*, núcleos dos antecedentes dos pronomes relativos, mas antes os dos sujeitos nulos:

(36)

a. [A *rapariga* que [-] **vês** ali] é minha vizinha.
b. [Os *livros* que [-] **comprei** no Natal] não foram caros.

(Brito e Raposo, 2013: 1103; com grifos nossos)

Tendo em vista os dados a serem analisados, interessa ressaltar o papel e a flexão dos pronomes relativos, constituintes que, segundo Veloso (2013: 2070), introduzem as orações relativas, exercendo uma função gramatical dentro delas. De acordo com essa autora, através desses pronomes, "há uma retoma pronominal do antecedente, através dos traços semânticos correspondentes do pronome, os quais codificam noções equivalentes ou hiperónimas do antecedente" (Veloso, 2013: 2076).

Quanto à morfologia, há os pronomes variáveis (*o qual*, *a qual*, *os quais*, *as quais*, *cujo*(*s*), *cuja*(*s*), *quanto*(*s*), *quanta*(*s*)) e os pronomes invariáveis (*que*, *o que*, *quem*, *onde*, *como*, *quando*, *quanto*) (Veloso, 2013: 2078). O pronome relativo invariável *que*, apresentado em (37a), não se flexiona, ainda que partilhe necessariamente os traços do seu antecedente. Já nos os pronomes relativos variáveis *os quais* (37b) e *as quais* (37c) concordam em número e gênero com os antecedentes retomados (*dados* e *duas figuras políticas*).

(37)

a. Itamar admite *a hipótese* de **que** pode ter se apaixonado....
b. ...usou *dados* d**os quais** discordamos.
c. ... citou os dois líderes como *duas figuras políticas* pel**as quais** ele nutria...

(Corpus: CETENFolha)

5.1.2.1.2 Verbos não-finitos: o caso do infinitivo flexionado

Estão incluídas nas formas verbais não-finitas o infinitivo, o gerúndio e o particípio. Em português, diferente da maioria das línguas, os verbos no infinitivo podem assumir uma forma flexionada ou não flexionada. Por isso, será dada uma atenção especial aos casos de infinitivo flexionado na presente seção, uma vez que apresentam também marcas de pessoa e número, em concordância com o sujeito.

Segundo Barbosa e Raposo (2013: 1904), "na medida em que incluem inequivocamente um sujeito, as construções de infinitivo flexionado têm claramente um

estatuto oracional, semelhante ao das orações finitas". Esse sujeito pode ser realizado foneticamente, como em (38a), ou pode ser um sujeito nulo, como em (38b).

(38)

a. ...este ano foi a vez d*os holandeses* **invadirem** Paris...
b. Penso em lugar de [-] **usarmos** o voto para [-] **estimularmos** o eleitor...

(Corpus: CETENFolha)

Quanto à flexão desses verbos em perífrases verbais finitas, Raposo (2013: 1226) aponta que apenas o primeiro verbo da sequência, ou seja, o verbo auxiliar mais à esquerda, contém as marcas flexionais de tempo, modo, aspecto, e de pessoa e número, nestes dois casos em concordância com o sujeito. Tendo em vista essa característica, "os verbos auxiliares não se combinam com um verbo no infinitivo flexionado" (Raposo, 2013: 1240). A agramaticalidade dos seguintes excertos em PE exemplifica esse impedimento:

(39)

a. *As bombas *vão* ***explodirem***.
b. *Eles *podem* ***jogarem*** na lotaria.

Ainda relativamente à flexão, "as desinências do infinitivo flexionado são acrescentadas ao morfema *–r* do infinitivo" (Barbosa e Raposo, 2013: 1904). A tabela 1 abaixo, extraída de Duarte et al. (2016: 453), exemplifica a conjugação em PE desse tipo de verbo nas três pessoas gramaticais (singular e plural) com verbos regulares:

	Número	
Pessoa	**Singular**	**Plural**
1.ª	Cantar, comer, partir	Cantar**mos**, comer**mos**, partir**mos**
2.ª	Cantar**es**, comer**es**, partir**es**	Cantar**des**, comer**des**, partir**des**
3.ª	Cantar, comer, partir	Cantar**em**, comer**em**, partir**em**

Tabela 1 – Conjugação do infinitivo flexionado

(extraída de Duarte et al., 2016: 454)

As próprias autoras ressaltam que a forma de 2ª pessoa do plural, correspondente ao pronome na forma nominativa "*vós*", caiu em desuso no português. Além disso, ao observar a tabela se nota que "no singular, apenas a forma de 2ª pessoa apresenta

morfemas explíctios (*-es*), que permitem distingui-la da forma do infinitivo não flexionado, sendo as restantes homónimas" (Duarte et al., 2016: 454).

Segundo Barbosa e Raposo (2013: 1904), quando o sujeito é um SN (40a) ou um pronome (40b) foneticamente realizado, o infinitivo deve assumir sua configuração flexionada, caso contrário incorre em agramaticalidade como em (40c).

(40)

a. Eu lamento *os meus pais* **terem** insultado a professora.
b. Eu lamento *tu* **teres** insultado a professora.
c. *Eu lamento *os meus pais*/*tu* **ter** insultado a professora

(Barbosa e Raposo, 2013: 1904)

No caso de orações com sujeito não realizado foneticamente, verbos como *lamentar* possuem a propriedade de selecionar ambas as formas flexionada (41a) e não flexionada (41b).

(41)

a. Eu lamento [-] **terem**/**termos** insultado a professora
b. Os alunos lamentam ter insultado a professora.

(Barbosa e Raposo, 2013: 1904)

O infinitivo flexionado possui uma distribuição bastante restrita, ocorrendo tipicamente em contextos de frase subordinada (Duarte et al., 2016: 455). Segundo essas autoras, "em contextos de frases-raiz, só em enunciados com valor ilocutório avaliativo e expressivo (exclamativas e optativas) é possível encontrar infinitivo flexionado" (Duarte et al., 2016: 455). Esse tipo de enunciado é exemplificado nos excertos seguintes excertos:

(42)

a. **Viver*mos*** sem medo! Quem nos dera!
b. **Poder*es*** fazer o que te apetece sem ninguém a mandar em ti!

(Duarte et al., 2016: 455)

Segundo Duarte et al. (2012: 219), essa forma verbal usualmente ocorre em sujeitos frásicos (43a), em orações adjuntas (43b) e em completivas de objeto selecionadas por certos verbos (43c). No exemplo (43c), é possível observar que a forma flexionada pode ser selecionada por verbos epistêmicos, declarativos e factivos, respectivamente, como também assinalou anteriormente Raposo (1987: 87).

(43)

a. Surpreendeu a Maria ***termos*** *chegado atrasados à reunião*.
b. A mãe arrumou a casa *antes de os meninos* ***chegarem***.
c. Eles pensam/afirmam/lamentam ***termos*** *chegado atrasados*.

(Duarte et al., 2012: 219)

Também é possível a forma flexionada em complementos selecionados por certos nomes e adjetivos, como exemplificam o nome *desejo* em (44a) e o adjetivo *ansiosos* (44b).

(44)

a. Eles sempre confessaram o desejo *de* ***emigrarem***.
b. Eles estão ansiosos *por* ***irem*** *de férias.*

(Duarte et al., 2016: 459)

No caso de complementos subcategorizados por predicados volitivos, por um lado, em PE[13], de modo geral, não é possível ocorrer infinitivo flexionado (Raposo, 1987: 87), como é possível verificar na agramaticalidade de (45a). Por outro lado, pesquisadores apontam que em PB, é possível o infinitivo flexionado com verbos volitivos (cf. Da Luz, 1998; e Groothuis, 2015), como é possível observar em (45b).

(45)

a. *Eu *desejava* [os deputados **terem** trabalhado mais].

(Raposo, 1987: 88)

b. O presidente *preferiu* [se **reunirem** às 6h].

(Groothuis, 2015: 18)

É possível encontrar exemplos do *corpus* consultado que confirmam o ocorrência de infinitivo flexionado em subordinadas com função de sujeito frásico (46a) e adjunto (46b). Também foram encontrados exemplos de infinitivo flexionado em complementos oracionais selecionados por nome (*tentativa* em (47a)) e adjetivo (*capazes* em (47b)). Quanto aos verbos, não foram encontrados exemplos de infinitivo flexionado em completiva oracional de verbo volitivo, mas foram encontrados em completivas selecionadas por verbo declarativo (*afirmam* em (48a)), espistêmico (*pensei* em (48b)) e causativo (*mandei* em (48c)).

(46)

a. ...parece difícil ***construirmos*** *uma teoria da ação social* quando...
b. *Após o* ***libertarem***, os ladrões foram seguidos...

[13] Em PE também se encontram atestadas ocorrências de infinitivo flexionado com alguns verbos volitivos, mas que não são padrão (cf. Gonçalves et al., 2014). Esses casos não serão discutidos aqui, tendo em vista os objetivos da presente pesquisa.

(47)

a. ...consideram uma tentativa *de* ***desmerecer*** *a conquista do tetracampeonato*.
b. ... o dos partidos capazes *de* ***assegurarem*** *a vitória* e a maioria parlamentar...

(48)

a. Cientistas do Canadá, EUA e Cingapura afirmam ***terem*** *produzido um salmão-monstro*...
b. Coisas que sempre pensei ***terem*** *sido esquecidas* continuavam a remoer a intimidade do « sr. Diretas».
c. Foi por isso que eu mandei ***ligarem*** *ao hospital* para checar.

(Corpus: CETENFolha)

Ao observar o funcionamento da flexão do infinitivo em textos orais do PB, Monteiro (1996) aponta as dificuldades encontradas por gramáticos na "elaboração de regras para o uso do infinitivo flexionado em português, de tal forma que não é difícil encontrar contradições ou divergências entre o que se prescreve e o que se pratica" (Monteiro, 1996: 62).

Partindo de textos orais em PB, esse autor aponta algumas dessas contradições entre o que é prescrito e o que é praticado. Primeiramente, "realmente em muitos enunciados o infinitivo se flexiona quando apresenta um sujeito próprio" (Monteiro, 1996: 62), como em (49). Entretanto, "o infinitivo muitas vezes aparece inflexionado, mesmo quando apresenta sujeito próprio" (Monteiro, 1996: 63), como em (50).

(49)

a. ...a coisa d*os pais* **saberem** que as filhas...
b. ...para *vocês* **entenderem** bem

(Monteiro, 1996: 63)

(50)

a. isso implica em *índices de evaporação*$_i$ capazes de [-]$_i$ **salinizar** qualquer terra irrigada
b. ela está deixando primeiro esfriar mais né... **passar** mais *essas festas natalinas*

(Monteiro, 1996: 63)

Além disso, geralmente também é estabelecido que o infinitivo "não se flexiona quando seu sujeito é idêntico ao da oração principal" (Monteiro, 1996: 63), como em (51). Porém, é possível encontrar exemplos em que "ocorre a flexão do infinitivo, mesmo quando este não apresenta sujeito próprio" (Monteiro, 1996: 63), como nos excertos em (52).

(51)

a. *eles* corriam pra cima da gente pra **vender** as fitinhas
b. *nós* demoramos muito a **aprender** a linguagem do cinema

(Monteiro, 1996: 63)

(52)

a. *os jornalistas* todos os dias são muito convidados e convocados a ganhar propinas pra **falarem** bem ou mal...
b. *eles* não vão lá só pra ir se **apresentarem** e de repente...

(Monteiro, 1996: 63)

A partir da observação dessa diversidade no emprego do infinitivo flexionado, esse autor conclui que

> "[se] a flexão do infinitivo não é disciplinada por uma regra categórica, trata-se evidentemente de um fenômeno de variação. E, dessa forma, poderá não ser aleatória ou motivada apenas por elementos de natureza subjetiva ou estilística, como pretenderam alguns autores, mas, bem ao contrário, ser uma variação condicionada em função do ambiente lingüístico em que ocorre" (Monteiro, 1996: 65).

Os dados analisados por Monteiro (1996) são textos orais, mas também é possível observar essa variação em textos escritos. Para exemplificar, a análise do infinitivo flexionado em textos jornalísticos escritos em PB feita por Cabral (2006) demonstrou que certos fatores como, para citar alguns, a posição da oração da oração infinitiva, a distância entre o verbo da oração principal e da oração infinitiva, a presença de pronome apassivador ou a presença de predicativo plural podem favorecer a ocorrência da flexão.

Outras pesquisas acerca da variação no emprego do infinitivo no português escrito confirmam que o uso da flexão não é meramente aleatório, mas sim determinado por diversos fatores linguísticos, como a relação de "controle" entre o sujeito do infinitivo flexionado e os argumentos da oração principal, a dependência temporal entre a oração principal e a oração infinitiva e a orientação temporal do verbo superior (cf. Barbosa e Raposo, 2013; Duarte et al., 2012; Duarte et al., 2016).

Segundo Barbosa e Raposo (2013: 1941), nas construções de controle, a referência do sujeito da oração subordinada de infinitivo não flexionado é controlada por um dos argumentos da oração superior, constituindo-se controle obrigatório. No excerto (53a), o sujeito nulo do verbo infinitivo não flexionado *examinar* é controlado pelo SN sujeito da oração principal *os médicos*.

(53)

a. *Os médicos*$_i$ desejam [[-]$_i$ **examinar** a Ana]

(Barbosa e Raposo, 2013: 1941)

Nas construções de controle obrigatório, o antecedente do sujeito implícito da oração infinitiva pode ser o sujeito, o objeto direto ou o objeto indireto da oração principal (Barbosa e Raposo, 2013: 1942). Nos excertos apresentados a seguir, foram assinalados os verbos infinitivos e os antecedentes do sujeito nulo, respectivamente o sujeito (54a), o complemento direto (54b) ou o complemento indireto (54c) da oração principal. Nos dois últimos exemplos (54b) e (54c) os falantes do português também admitem o infinitivo flexionado.

(54)

a. *Os meus pais*$_i$ pensam [-]$_i$ **comprar** um automóvel novo.
b. Incentivámos *as crianças*$_i$ a [-]$_i$ **brincar** umas com as outras.
c. A criança pediu *aos pais*$_i$ para [-]$_i$ lhe **comprar** uma bicicleta.

(Barbosa e Raposo, 2013: 1942-1944)

Duarte et al. (2012, 2016), mostram que, nos contextos em que se verifica controle obrigatório de sujeito, o uso do infinitivo flexionado é restringido por "uma componente semântica (o requisito de tempo independente da oração infinitiva) e uma componente lexical (um verbo superior que selecione uma orientação temporal não especificada do complemento infinitivo)" (Duarte et al., 2016: 463).

Outros contextos em que ocorre infinitivo flexionado também são apresentados na pesquisa de Duarte et al. (2016). Porém, tendo em vista os objetivos do presente relatório, não serão aqui apresentados. Segue-se uma síntese fornecida pelas autoras acerca da distribuição do infinitivo flexionado, casos em que se verifica concordância com o sujeto, nulo ou realizado:

> "(i) Contextos em que o infinitivo flexionado tem uma distribuição livre: a) contextos que não envolvem controlo obrigatório do sujeito (...); b) com verbos de elevação de sujeito para objeto, com verbos causativos, percetivos e com o verbo de permissão *deixar*; c) na construção de infinitivo preposicionado. (ii) Contextos em que o infinitivo flexionado tem uma distribuição restrita: a) contextos de complementação selecionados por verbos de controlo de sujeito (...); b) contextos de controlo de objeto (obrigatório ou preferencial) (...) c) com verbos de elevação de sujeito para sujeito, o infinitivo flexionado pode ocorrer, não se registrando elevação" (Duarte et al., 2016: 477-478).

5.1.2.2 Concordância Sujeito-Predicativo do Sujeito

Diferente das predicações de base verbal, nas predicações de base adjetival e nominal, não são os verbos copulativos "que veiculam o conteúdo fundamental daquilo que se diz sobre o sujeito e, concomitantemente, a natureza da situação descrita" (Raposo, 2013: 1286). Esse conteúdo fundamental é da responsabilidade do nome, no caso da predicação nominal, ou do adjetivo, no caso da predicação adjetival. Diante disso, a base semântica da oração copulativa é o nome ou o adjetivo, mas o núcleo gramatical dos sintagmas verbais dos quais fazem parte esses predicadores é o verbo copulativo (Raposo, 2013: 1290). Nos excertos a seguir, o nome *trabalhadores* (55a) e o adjetivo *gratuita* (55b) são os predicadores das orações e são núcleos gramaticais do SV os verbos copulativos *são* (55a) e *é* (55b):

(55)

a. *Os italianos* também **são trabalhadores**.
b. *A participação* **é gratuita**.

Como é possível observar nos exemplos apresentados acima, sendo o núcleo gramatical da oração copulativa, o verbo copulativo concorda em pessoa e número com o sujeito da oração (Raposo, 2013: 1291). Nas predicações de base adjetival e nominal, o nome e o adjetivo que são núcleos do constituinte predicativo, também devem concordar em gênero e número com o sujeito da oração[14]. Em (55a) e (55b), ambos os predicativos flexionam em conformidade com os traços dos sujeitos *os italianos* (55a) e *a participação* (55b).

Também com o predicativo é possível verificar a ocorrência de variação na concordância em português, principalmente na língua oral, como apontam Scherre e Naro (1991). Apesar disso, não serão dados detalhes acerca desse tipo de variação tendo em vista os objetivos da presente subseção.

5.1.3 O papel da concordância na coesão textual

Segundo Mendes (2013), a coesão auxilia na conexão entre a informação já apresentada pelo falante e a informação nova, trazida em cada uma das frases que vão se sucedendo. Possui particular importância para a análise dos dados anotados e

[14] Há casos em que predicadores nominais podem não concordar com o sujeito. São casos em que a expressão nominal for qualitativa, como no exemplo (i):

(i) As intervenções dele **são** *um perigo*. (Duarte, 2003: 291).

analisados nesta pesquisa a coesão referencial, que auxilia na identificação das entidades referidas no texto. Para efetuar essa coesão entre os elementos, pode-se utilizar operações de *foricidade*, que envolvem elementos dentro do texto, ou de *dêixis*, que apontam para elementos no contexto pragmático.

As expressões assinaladas nos excertos a seguir se referem a elementos que se encontram fora do texto, sendo a referência deles estabelecida no universo do discurso, através de operações de *dêixis*. A sentença apresentada em (56a) exemplifica um SN cujo conteúdo descritivo auxilia na identificação do referente. No caso da sentença (56b), o pronome demonstrativo auxilia na identificação da entidade no contexto de enunciação.

(56)

a. Comprei ontem *a cômoda que pertencia à Maria Antonieta.*
b. (O falante, apontando para um livro) *Isto* custou 20 euros.

(Mendes, 2013: 1702)

Os pronomes *o* e *lhes* apontados em (57a) e (57b) se referem a entidades já citadas no enunciado e são dependentes de outra expressão referencial, criando uma relação anafórica dentro do texto. Os excertos apresentados também mostram que os antecedentes dos elementos anafóricos podem estar dentro da mesma frase (57a) ou em frase anterior (57b). De acordo com Lobo (2013) esses elementos são chamados correferentes, pois se referem a mesma entidade, formando assim uma cadeia referencial (Lobo, 2013: 2180).

(57)

a. Encontrei *o formulário* e entreguei-**o** à funcionária.
b. *Os vizinhos* ajudaram-me. Agradeço-**lhes** imenso.

(Mendes, 2013: 1702)

5.1.3.1 Relações de foricidade

Ao falar sobre foricidade, Castilho (2010) aponta que ela pode ser feita a partir da anáfora, na qual itens lexicais trazem novamente noções já identificada anteriormente, ou a partir da catáfora, em que a noção será identificada posteriormente no texto (Castilho, 2010: 125). Em PB, a anáfora é o caso mais comum desse tipo de retoma. O pronome *isso* em (58a) ilustra um caso de catáfora, em que seu referente será apresentado após os dois-pontos. Em (58b) o mesmo pronome funciona como anáfora, pois o seu referente já foi apresentado anteriormente no texto.

(58)

a. Foi **isso** que eu disse à «IstoÉ»: toda vez que tiver um troço desses eu dou um cacete.
b. «Só quero um minuto na TV, **isso** basta.

(Corpus: CETENFolha)

Para garantir a eficácia da retoma anafórica, ou seja, "para que possa haver relações de identidade referencial entre duas expressões nominais, é necessário que elas possuam especificações de gênero, número e pessoa que sejam compatíveis" (Lobo, 2013: 2186). Por isso a concordância verbal e nominal é essencial na construção de cadeias anafóricas. Para exemplificar, Lobo (2013) aponta que a substituição do pronome *os* (59a) pela forma feminina *as* (59b) nos exemplos a seguir exclui a leitura do pronome como correferente com o sujeito da oração principal, pois em (59b) não houve a partilha de traços entre o SN e o pronome:

(59)

a. *Os alunos* disseram que os professores **os** tinham magoado.
b. Os alunos disseram que os professores **as** tinham magoado.

(Lobo, 2013: 2186)

No caso da concordância verbal, a partilha de traços de pessoa e número auxilia na estruturação de cadeias anafóricas. A importância desse tipo de concordância fica evidente no caso dos constituintes nulos: é importante que os traços de pessoa e número concordem com os do seu antecedente para a correta interpretação dessas expressões. No excerto (60a), a partir da flexão verbal na 3ª pessoa do plural, é possível perceber que o SN "*as pessoas*" é correferente com o sujeito nulo dos verbos "*tiram*" e "*puderem*". Já no exemplo (60b), tendo em vista a flexão dos verbos assinalados estar na 3ª pessoa do singular, os sujeito nulos ligados a esses verbos são correferentes.

(60)

a. Não sei por que **as pessoas**$_i$ não aceitam isso como uma fantasia e [-]$_i$ **tiram** daí o prazer que [-]$_i$ **puderem**.
b. [-]$_i$ Não **sabe** por que as pessoas não aceitam isso como uma fantasia e [-]$_i$ **tira** daí o prazer que [-]$_i$ **pode**.

(Corpus: CETENFolha, com adaptações e grifos nossos)

5.1.3.2 Relações de dêixis

Segundo Castilho (2010), a referência de termos ou expressões dêiticas está no discurso, ou seja, na situação concreta que envolve os falantes e não apenas no

significado das palavras (Castilho, 2010: 123). O sentido desses elementos se organiza na esfera pragmática e depende da situação de fala em que foram veiculados.

Para a análise dos dados anotados na ferramenta da Unbabel, é interessante a dêixis pessoal, que está relacionada com a coesão referencial. Segundo Raposo (2013), esse tipo de dêixis corresponde à referência aos participantes no ato da enunciação, sendo possível identificar os participantes (falante e ouvinte) e outras entidades presentes no contexto (Raposo, 2013: 395). Os pronomes pessoais de primeira e segunda pessoa (*eu*, *tu*, *nós*, *você, vocês, a gente*), bem como as formas possessivas (*meu*, *teu*, *seu, nosso*, *vosso*) são centrais nessa identificação. Esses tipos de pronomes são exemplificados a seguir:

(61)

a. Na segunda-feira, troquei **seu** cheque pelo **meu**», disse.
b. Para **você**, leitor, não poderia ser melhor.

(Corpus: CETENFolha)

A flexão verbal também é apontada por Raposo (2013: 395) como um dos instrumentos centrais na referência dêitica nos casos de elipse ou de sujeito não realizado. Em (62a), graças a flexão verbal, é possível distinguir o sujeito de ambos os verbos *pode* e *aceitamos*: o primeiro sujeito nulo é correferente com "*o PT de lá*", já o segundo sujeito nulo é dêitico, remetendo ao próprio falante, pois possui traços de 1ª pessoa do plural.

(62)

a. «Se *o PT de lá*$_i$ quiser, [-]$_i$ **pode** mandar o dinheiro que [-] aceitamos».

(Corpus: CETENFolha, com adaptações nossas)

Para a análise dos dados fornecidos, é interessante notar os casos de concordância em que as informações acerca dos traços de gênero e número das entidades não estão explícitas no texto, mas são estabelecidas pragmaticamente. Nesses casos, para realizar a concordância, é necessário ter conhecimentos acerca do contexto específico em que se deu o enunciado, como ocorre nos casos de elementos dêiticos.

(63)

a. [-] Relata que outro dia foi **obrigada** a recusar comida a uma mulher que vive nas ruas do seu bairro.

(Corpus: CETENFolha)

Em (63a), por exemplo, tendo em vista a omissão do sujeito de "*relatar*", é necessário ter conhecimentos acerca do contexto de enunciação para saber o gênero do sujeito e determinar se o particípio passado passivo "*obrigada*" se flexiona no masculino ou no feminino.

5.2 Questões sobre a ordem de palavras em PB

Nesta seção serão observados fenômenos ligados à ordem de palavras em PB. Para facilitar a discussão, os tópicos descritos foram divididos em duas subseções: (5.2.1) trata da ordem dos constituintes em PB a nível frásico e (5.2.2) observa a ordem dos elementos dentro do SN em PB.

5.2.1 Ordem dos constituintes da frase

Tendo em vista os fenômenos presentes nos dados fornecidos e a extensão da presente pesquisa, preferiu-se dar atenção especial ao funcionamento dos clíticos e dos advérbios, tendo em vista as particularidades do posicionamento desses elementos dentro das frases em PB. A presente seção se organiza da seguinte forma: na seção 5.2.1.1 é feita uma apresentação geral da ordem entre o sujeito, o verbo e o complemento; na seção 5.2.1.2 são observadas as regras relativas à posição dos clíticos em PB; e na seção 5.2.1.3 são apresentadas as possíveis posições do advérbio na frase.

5.2.1.1 Questões iniciais

Na ordem não marcada de sentenças declarativas em português, o argumento externo do núcleo é gerado fora do SV e se posiciona à esquerda do verbo enquanto os argumentos internos (ou complementos) são gerados dentro do SV e se posicionam à direita do verbo (Raposo, 2013: 366). Essa ordem não marcada é geralmente representada por SVO (Sujeito - Verbo - Objeto). Para exemplificar, em (64a), *Asilados* é o sujeito da oração (S), *inauguraram* é o verbo (V), núcleo do SV, e *embaixada da Iugoslávia* é o complemento (O). A mesma ordem é exibida em (64b).

(64)

a. [Asilados] *inauguraram* [embaixada da Iugoslávia] SVO
b. [Cinco dos 27 ministros] *não voltaram* [a Brasília] SVO

(Corpus: CETENFolha)

Note-se, porém, que é possível alterar a ordem básica, como se verifica nos exemplos (65), retirados do texto de Martins e Costa (2016):

(65)

a. [Esse tipo de notícia] pouco *interessa* [ao cidadão comum]. SVO
b. [Ao cidadão comum] pouco *interessa* [esse tipo de notícias]. OVS
c. Pouco *interessa* [ao cidadão comum] [esse tipo de notícias]. VOS
d. Pouco *interessa* [esse tipo de notícias] [ao cidadão comum]. VSO
e. [Ao cidadão comum], [esse tipo de notícias] pouco *interessa*. OSV
f. [Esse tipo de notícias], [ao cidadão comum] pouco *interessa*. SOV

(Martins e Costa, 2016: 372, com grifos nossos)

Porém, como é afirmado por esses autores, a ordem dos elementos frásicos não é "livre" em português, pois ela depende de diversos fatores sintáticos, semânticos e pragmáticos. Além disso, algumas posições podem gerar sequências agramaticais, como exemplifica o excerto (66b) a seguir, retirado do texto desses autores, que apresenta uma ordem OSV não adequada. Ainda segundo esses autores, o contraste entre as frases (65e), gramatical na ordem OSV, e (66b), agramatical nessa mesma ordem OSV, reside na diferença entre entre as estruturas informacionais das duas frases[15]: em (65e), a presença do modificador "pouco" antes do verbo torna o comentário expresso pelo predicado "pouco interessa" informacionalmente relevante, já em (66b) o predicado "estarão" não possui uma quantidade satisfatória de conteúdo semântico e relevância informacional.

(66)

a. [Parlamentares como Mariana Mortágua, do BE, ou João Galamba, do PS,] *estarão* [na Comissão Parlamentar de Inquérito]. SVO
b. *[Na Comissão Parlamentar de Inquérito], [parlamentares como Mariana Mortágua, do BE, ou João Galamba, do PS,] *estarão*. OSV

(Martins e Costa, 2016: 373, com grifos nossos)

Apesar de a ordenação SVO também ter sido indicada como a ordem não marcada do PE e PB por Kato e Martins (2016), essas pesquisadoras explicitam as diferenças entre a ordem Verbo-Sujeito e Sujeito-Verbo nessas duas variantes, apontando casos, como os ilustrados em (67), em que a estrutura com sujeito pós-verbal é considerada a ordem mais natural. Isso demonstra a grande variação que pode ocorrer na ordem dos elementos dentro do SV. Entretanto, tendo em vista a extensão da presente pesquisa, decidiu-se não alongar essa questão.

(67)

a. *Chegou* [a primavera];
b. *Passaram* [poucos alunos] no exame;

[15] Para mais informações acerca das estratégias gramaticais ligadas à alteração da ordem básica, dos fatores que propiciam ou reprimem as diferentes posições dos constituintes frásicos, bem como das diferentes interpretações resultantes dessas alterações de ordem, cf. Martins e Costa (2016) e Kato e Martins (2016), entre outros.

c. *Viajou* comigo [um cantor de rock].

(Kato e Martins, 2016: 24, com grifos nossos)

Acerca da ordem dos complementos, o objeto indireto se coloca tipicamente após o objeto direto, como nos exemplos em (68), mas também é possível encontrar o objeto indireto antecendendo o complemento direto como nos casos em (69).

(68)

a. ...Rincón *deu* [uma entrevista] [a uma rádio colombiana]...
b. ...os bicheiros *emprestariam* [sua estrutura] [aos traficantes internacionais]...
c. O congresso *oferece* [sugestões] [às famílias e comunidades]...

(Corpus: CETENFolha)

(69)

a. Vamos *dar* [a eles] [um presente de grego] nesse dia...
b. ...*empresta* [aos produtores] [US$ 11 milhões]...
c. *Oferece* [ao usuário] [duas alternativas]...

(Corpus: CETENFolha)

5.2.1.2 Ordem de complementos verbais: o caso particular dos clíticos

Nesta seção será abordada a ordem das formas átonas acusativas e dativas dos pronomes pessoais, buscando-se compreender as regras que norteiam o posicionamento desses elementos em PB. As variedades brasileira e europeia do português se diferenciam muito quanto ao uso dos clíticos, sendo a posição dessas formas pronominais átonas um dos aspectos que mais distingue as duas variedades (cf. Luís e Kaiser 2016; Duarte 2015; Duarte 2013; Mattos e Silva 2013; Galves 2001; e Pagotto 1992). Os próximos parágrafos pretendem fornecer uma comparação sucinta dos clíticos em ambas as variedades e explicitar a evolução do paradigma pronominal em PB que influencia enormemente o uso dos clíticos nessa variedade.

Segundo Duarte, todos os pronomes clíticos "exigem um hospedeiro verbal, que se traduz num requisito de adjacência entre o pronome clítico e uma forma verbal, finita ou não finita" (2013: 847). As formas átonas dos pronomes podem se ligar ao verbo se posicionando à sua esquerda (próclise), à sua direita (ênclise) ou no interior da própria forma verbal no condicional ou no futuro (mesóclise[16]). As seguintes sentenças em PE exemplificam essas posições, respectivamente:

[16] Interessante ressaltar que, segundo Duarte, no PE moderno a mesóclise é uma posição sobrevivente dos traços de "uma gramática antiga, claramente em desaparecimento" (2013: 865).

(70)

a. Não ***as*** *vi* ontem à noite (próclise)
b. *Vi*-***as*** ontem à noite (ênclise)
c. *Vê*-***las***-*ei* à noite / *Vê*-***las***-*ia* à noite (mesóclise)

(Raposo, 2013: 905)

Como é possível verificar na tabela 2 a seguir, feita com base em Raposo (2013: 902), os clíticos em PE têm variantes em número nas três pessoas gramaticais, com exceção de *se*, e variantes em gênero nas formas *o* e *a* (Raposo, 2013: 903).

	Formas Átonas (Clíticas)	
	Complemento Direto (Acusativo)	**Complemento Indireto (Dativo)**
1sg	me	me
1pl	nos	nos
	nos, se	nos, se
2sg	te	te
	o, a, se	lhe
2pl	vos	vos
	os, as, vos, se	lhes, vos
3sg	o, a, se	lhe
3pl	os, as, se	lhes

Tabela 2 – Formas átonas do PE

(Raposo, 2013: 902)

Galves (2001: 125) aponta uma mudança no paradigma pronominal do PB, que o diferencia do PE, em particular no que diz respeito aos clíticos. A tabela 3 apresentada a seguir evidencia essa transformação.

	Nominativo	Acusativo	Dativo	Oblíquo
	1. eu	me	me	mim
Singular	2. *tu/você	te/você/lhe	lhe/a você	ti/você
	3. ele (ela)	ele(ela)/o(a)	a ele(ela)	ele(ela)
Plural	1. nós/a gente	nos/a gente	nos/a gente	nós/a gente
	2. vocês	vocês	a vocês	vocês
	3. eles(elas)	eles(elas)/os(as))	a eles(elas)	eles(elas)
* Uso dialetal				

Tabela 3 – Paradigma Pronominal do PB

(extraída de Galves, 2001: 129)

Ao observar as formas acusativas apresentadas nessa tabela, outras características ficam evidentes: as formas não clíticas *você*(*s*), *a gente*, *ele*(*s*) e *ela*(*s*) também podem ser utilizadas em PB em contextos em que o PE recorre a formas clíticas. Além disso, o clítico *lhe*, com uso dativo em PE , alterna com a forma preposicional em que o pronome pleno é introduzido pela preposição "*a*".Essas particularidades do PB também foram ressaltadas por Luís e Kaiser (2016: 225). Repetimos aqui os exemplos desses autores que demonstram a substituição de uma forma clítica por uma forma não clítica: *te* por *você* (71a), *nos* por *a gente* (71b); e *lhe* por *a ela* (71c):

(71)

a. Não **te** chamei. / Não chamei *você*.
b. A Maria viu-**nos** na praia. / A Maria viu *a gente* na praia.
c. O João não **lhe** perguntou. / O João não perguntou *a ela*.

(Luís e Kaiser, 2016: 225)

Estes aspectos permitem afirmar que os clíticos estão em processo de perda no PB, como afirma Mattos e Silva (2013): "os pronomes clíticos com a função de complemento, sobretudo os de terceira pessoa (*o*(*s*) e (*a*(*s*)), estão em perda no português do Brasil (...) em seu lugar ocorrem muito mais frequentemente o sintagma nominal pleno e o chamado 'ele acusativo' ou pronome forte" (Mattos e Silva, 2013: 152).

Acerca do pronome *lhe*, essa autora afirma que "é de notar a perda acentuada da sua função como dativo, (isto é, como complemento indireto), e o seu uso crescente, em certas variedades dialetais, como acusativo (ou seja, como complemento direto) na 2ª pessoa" (Mattos e Silva, 2013: 152). Nos excertos extraídos do texto dessa autora, é possível observar o uso acusativo de *lhe* em PB (72a), caso em que o clítico *o* seria utilizado em PE (72b):

(72)

a. PB: *Você* gosta mesmo de golfe! Eu **lhe** vejo sempre no clube.
b. PE: Eu vejo-**o** sempre no clube.

(Mattos e Silva, 2013: 152)

É possível concluir que os clíticos sejam uma questão muito debatida na descrição gramatical do PB, tendo em vista essa variação e mudança no uso dos clíticos. À luz do que foi sugerido por Duarte (2015), procura-se na presente descrição considerar as pesquisas linguísticas mais recentes acerca do uso efetivo dos clíticos nos textos escritos do PB, para tentar compreender as escolhas feitas pelos anotadores durante a anotação

dos dados analisados neste trabalho. Tendo isso em consideração, serão citados exemplos retirados de pesquisas feitas nessa área juntamente com excertos do *corpus* consultado que confirmem as características do PB expostas por essas pesquisas.

5.2.1.2.1 Alguns estudos acerca da tendência proclítica em PB

Diversos estudos (cf. Faria 2017; Pagotto 1992; Silva 2017) demonstram que a próclise é a posição mais natural em PB. Galves (2001) ressalta que em PB "a ênclise não é totalmente ausente no uso, mas, ao contrário do PE, ela é fortemente ligada à forma do clítico" (Galves, 2001: 133). Na escrita monitorada na variante culta "existe, no entanto, uma tendência crescente para a ênclise em contextos em que até o português europeu sempre usou a próclise (...)" (Mattos e Silva, 2013: 153). Essa predisposição à hipercorreção da próclise também foi assinalada por Duarte (2015). Segundo a autora, no Brasil, diversos escritores provenientes de uma geração mais madura usam a ênclise apesar dos atratores de próclise, pois há uma proliferação da ideia de que a colocação dos pronomes proclíticos é uma questão de informalidade repudiada na escrita ou na fala mais formal (Duarte, 2015: 26-28).

(73)

a. O Aterro não está degradado, degradados estão os interesses *que* degradam-**no**. (Artigo em jornal)
b. *Ou* assume-**se** a segregação explícita, ou promove-se a miscigenação social. (Artigo em jornal)
c. Vem aí uma nova pedalada, daquelas que passam despercebidas [*porque* são complicadas] e [tornam-**se** simples] quando aparecem como tungas. (Artigo em jornal)

(Duarte, 2015: 26)

Como diz a própria autora acerca dos exemplos citados acima, "a ênclise do clítico ao verbo parece ter passado a compor com ele um vocábulo fonológico de tal forma que não importa a presença dos famosos proclisadores ou atratores, expressos em (73a) e (73b) e oculto em (73c)" (2015: 27, com adaptação da numeração). Essa predisposição pode ser vista mesmo em exemplos de textos jornalísticos do CETENFolha, em que estão presentes elementos atratores de próclise como o elemento negativo "*nem*" (74a) e a conjunção "*porque*" (74b).

(74)

a. ...de que não irá embarcar *nem deixar-**se*** enredar na barganha por...
b. ...viu-se obrigado a desistir, *porque tornou-**se*** público que...

(Corpus: CETENFolha)

Luís e Kaiser (2016: 210-253) fizeram um paralelo entre a posição dos clíticos em PE e PB, demonstrando que as regras de colocação de ambas as variedades são diferentes. Uma dessas diferenças é a impossibilidade de os verbos em PB expressarem seus argumentos através de mais de um clítico. Assim sendo, formas como as frases em PE (75a) e (75b) não são possíveis em PB (Luís e Kaiser, 2016: 211).

(75)
a. Dá-se-lhe o remédio
b. Passou-se-me!

(Luís e Kaiser, 2016: 211)

Além disso, segundo esses autores, a posição típica dos clíticos em PB é pré-verbal (Luís e Kaiser, 2016: 223), mesmo nos casos em que a ênclise é exigida em PE como em contexto de frase simples com sujeito nulo (76a), de frase com um só verbo principal (76b) ou de frase afirmativa com verbo no imperativo (76c). Tendo em vista a posição dos clíticos em PB ser independente da presença de atratores de próclise, sentenças como (76d) são semelhantes às equivalentes em PE.

(76)
a. **Me** chamo Maria. / *Chamo-**me** Maria.
b. A médica **me** chamou.
c. **Me** chama! / *Chama-**me**!
d. O médico *não*/*já*/*nunca* **me** chamou.

(Luís e Kaiser, 2016: 223)

Também é possível verificar essas estruturas no *corpus* consultado, como exemplificam as seguintes estruturas com clítico na posição inicial da frase (77a), em frase com um só verbo principal e sujeito expresso (77b), em frase imperativa (77c) e em frase com atrator de próclise (77d):

(77)
a. **Me** sinto em condições.
b. Ela **me** dirigiu.
c. «**Me** larga, idiota! ", exclamava, indignado.
d. Eu *não* **me** considero nada.

(Corpus: CETENFolha)

Esses autores também afirmam que em sentenças com verbos (semi)auxiliares em PB, o pronome clítico se liga ao verbo temático como um proclítico (Luís e Kaiser, 2016: 223).

Os seguintes exemplos, retirados do texto dos autores, mostram sequências verbais com infinitivo (78a), particípio (78b) e gerúndio (78c), respectivamente.

(78)
a. A senhora poderia **me** *dizer* o seu nome?
b. Você tinha **me** *dito* que ficava.
c. Estou **te** *levando*...

(Luís e Kaiser, 2016: 223)

Essas mesmas estruturas de próclise em sequências verbais com verbos no infinitivo (79a), gerúndio (79b) e particípio (79c) também podem ser encontradas no *corpus* de PB escrito consultado:

(79)
a. ...eu *posso* ***te*** *dar* uma foto.
b. «Eu *fico* ***me*** *desdobrando*, pois moro...
c. ...que o sr. álvaro Dias *teria* ***me*** *chamado* de idiota.

(Corpus: CETENFolha)

Apesar dessa tendência generalizada do uso proclítico de pronomes observada em PB, há contextos em que a ênclise é favorecida. Luís e Kaiser (2016) citam tipos de ênclise em PB: "one involving the clitic *se* in root matrix clauses and another one the 3rd person clitic *o* / *a* in 'Aux Vinf' structures" (Luís e Kaiser, 2016: 224). Esses casos podem ser observados nos exemplos a seguir:

(80) Clítico "*se*" em sentenças com um só verbo principal
a. *Chegou*-**se** à conclusão.
b. *Parte*-**se** um ovo.

(81) Clíticos "*o/a*" em estruturas "auxiliar + infinitivo"
a. Não seria conveniente *mudá*-**lo**.
b. Você vai *levá*-**lo** a encontrar uma solução.

(Luís e Kaiser, 2016: 224)

De acordo com Luís e Kaiser (2016: 224), o fato de as estruturas em (80a) e (80b) serem impessoais explica a restrição da posição enclítica de "*se*". Além disso, a ênclise das formas "*o/a*" nos exemplos (81a) e (81b) parece ser determinada pela morfologia do verbo temático, que é não-finito. Segundo esses autores, também é possível observar o uso da ênclise em estruturas de verbos auxiliares com outros tipos de clíticos, mas a

ocorrência pós-verbal desses clíticos em PB parece ser um vestígio da norma culta escrita ensinada na escola (Luís e Kaiser, 2016: 224).

Há também casos em que a posição pós ou pré-verbal é opcional em PB (Luís e Kaiser, 2016: 225). Esses pesquisadores citam sentenças preposicionais com verbos não-finitos, como nos seguintes exemplos retirados do texto dos autores com próclise (82a) e ênclise (82b). Esse mesmo tipo de estrutura pode ser encontrada no *corpus* consultado, como é possível verificar em (83a) e (83b).

(82)

a. Estou aqui *para* **te** dizer que...
b. Estou aqui *para* dizer-**te** que...

(Luís e Kaiser, 2016: 224)

(83)

a. ...vou arrumar um negócio *para* ***te*** *distrair*.
b. ...em 1991, *por tratar-****se*** de monopólio,...

(Corpus: CETENFolha)

Também segundo Galves (2001: 133-134), a próclise em relação ao verbo principal é a regra geral em locuções verbais com o formato "auxiliar + particípio", "auxiliar + gerúndio" ou "verbo modal + infinitivo", como já foi possível verificar nos exemplos de Luís e Kaiser (2016) em (78) e nos excertos retirados do *corpus* consultado em (79). Essa autora aponta que encontrou vários exemplos em que mais de um termo se intercala entre o verbo auxiliar e o pronome clítico, o que comprova a ideia de que o clítico se liga ao verbo temático, como em (84a) e (84b) (Galves, 2011: 134).

(84)

a. Todos *podiam*, em termos industriais, **se** *desenvolver*.
b. Não *posso* no momento **lhe** *dar* uma resposta.

(Galves, 2001: 134)

Além disso, o clítico se mantém em próclise ao verbo pleno mesmo quando há uma negação, um complementizador ou certos advérbios tipicamente considerados atratores do clítico (Galves, 2001: 134). É possível verificar esse aspecto nos exemplos a seguir em que a locução verbal é precedida pelo advérbio *não* (85a), pelo complementizador *que* (85b), e pelo advérbio "atrator" *já* (85c):

(85)

a. Agora <u>não</u> *tinha* **me** *lembrado*.

b. Essas indústrias novas que *estão* **se** *implantando*.
c. Alguém já *podia* **me** *dizer*.

(Galves, 2001: 134)

No caso de frases com o infinitivo, o pronome se insere em próclise qualquer que seja a classe do verbo principal, mesmo em sentenças iniciadas por preposição (Galves, 2001: 134) como nos exemplos (86a) e (86b), retirados do *corpus* consultado.

(86)
a. ...dias de prazo *para* ***se*** *explicar* à Justiça...
b. ...o ânimo *de* ***me*** *convidar*, eu...

(Corpus: CETENFolha)

Galves (2001: 134) ressalta que os clíticos acusativos "*o/a*" possuem um comportamento diferente dos outros pronomes, pois mesmo com o verbo no infinitivo eles aparecem frequentemente enclíticos. A ênclise em frases com verbos infinitivos pode ser observada nos seguintes exemplos retirados do *corpus* consultado:

(87)
a. ...deixando a opção de *usá-****lo*** ao indivíduo.
b. ...mas vale *mencioná-****la*** aqui...

(Corpus: CETENFolha)

Esse comportamento diferenciado também pode ser observado com locuções verbais compostas por particípio, gerúndio ou infinitivo, em que há preferência pela ligação ao verbo (semi)auxiliar (flexionado). Galves (2001: 135) aponta que os falantes preferem que esse tipo de clítico "suba" até o verbo flexionado, preferindo a sentença representada em (88a) e não a representada em (88b):

(88)
a. Não **o** estava vendo.
b. *Não estava **o** vendo.

(Galves, 2001: 135)

Não é o foco da presente pesquisa entrar em detalhes no conceito de subida de clítico, por isso não alongaremos essa questão, mas é interessante comparar os exemplos (88) de Galves (2001: 135) com os seguintes excertos:

(89)

a. A Maria *podia* **te** *ajudar* nos trabalhos da escola
b. ??A Maria **te** *podia ajudar* nos trabalhos da escola

(Kanthack, 2002: 19)

Segundo Kanthack (2002: 113-114), há estranhamento em (89b), pois o PB não exibe com facilidade a subida do clítico (*clitic climbing*), em que um clítico se move da sentença mais baixa para a mais alta[17].

Em resumo, Galves conclui que "o clítico em PB é não apenas proclítico, mas também fortemente atraído (...) pelo verbo que lhe atribui sua função temática. O clítico acusativo, ao contrário, mostra uma tendência à ênclise e é mais atraído pelas formas verbais flexionadas" (Galves, 2001: 135).

Através do exame de pesquisas anteriores que já investigaram acerca do uso efetivo desses elementos pelos falantes brasileiros, Kanthack (2002) sugere uma descrição geral do comportamento dos clíticos em PB que considere ambas as formas dos clíticos e dos verbos. Para explicitar as propriedades dos clíticos em PB, Kanthack (2002: 113-118) divide os clíticos do PB em dois grupos: (1) *me*, *te*, *se*, *lhe*, *nos* (e variantes); (2) *o* (e variantes). Segundo ela, a posição dos clíticos não é homogênea e, dependendo do contexto, os clíticos do grupo "*o*" possuem um comportamento diferente.

Em sentenças com um só verbo finito, a posição mais frequente é a proclítica para os clíticos de primeiro (90a) e segundo (90c) grupo. (Kanthack, 2002: 115)

(90)

a. Ele **me** visitou no hospital.
b. ? Ele visitou-**me** no hospital.
c. Ele **a** visitou no hospital.
d. ? Ele visitou-**a** no hospital.

(Kanthack, 2002: 115)

Esse comportamento difere em sentenças onde há palavras com valor negativo à esquerda do verbo (Kanthack, 2002: 115). Nesses casos, os clíticos de ambos os grupos ocorrem na posição pré-verbal, como é possível observar nos exemplos a seguir com as palavras *nunca* (91a e 91b) e *não* (91c e 91d). Nesse aspecto, a ordem dos clíticos em PB é semelhante a de PE.

(91)

a. Ele *nunca* **me** encontrou na saída do colégio.
b. * Ele *nunca* encontrou-**me** na saída do colégio.

[17] Para mais informações acerca desse fenômeno, veja-se Duarte (2003: 857-860).

c. Ele *não* **o** encontrou na saída do colégio.
d. * Ele *não* encontrou-**o** na saída do colégio.

(Kanthack, 2002: 115)

Essa preferência pela posição proclítica em ambos os grupos também pode ser encontrada em estruturas subordinadas, como ilustram os seguintes exemplos da autora: são agramaticais as sentenças em que os clíticos de primeiro (92b) e segundo (92d) grupo são enclíticos.

(92)
a. A Maria disse que o João **te** visitou no hospital.
b. * A Maria disse que o João visitou-**te** no hospital.
c. A Maria disse que o João **a** visitou no hospital.
d. * A Maria disse que o João visitou-**a** no hospital.

(Kanthack, 2002: 116)

Também foram encontradas sentenças com o mesmo comportamento no *corpus* consultado. O clítico de primeiro grupo está na posição proclítica ao verbo finito em frase simples (93a) ou subordinada (93b). O clítico de segundo grupo também se encontra antes do verbo em frase simples (94a) ou subordinada (94b):

(93)
a. Eles **me** *olhavam* com desprezo, ódio...
b. Estive pensando outro dia que **me** *daria* muito bem...

(Corpus: CETENFolha)

(94)
a. Ela **o** *puxou* para dentro...
b. ...o governo federal **o** *avisou* sobre a possibilidade...

(Corpus: CETENFolha)

Já com um verbo não-finito, a distribuição dos clíticos não é generalizada, pois o clítico "*o*" possui outro comportamento (Kanthack, 2002: 116-117). Enquanto os clíticos do primeiro grupo são proclíticos com verbos no infinitivo (95a), os do segundo grupo são preferencialmente enclíticos (95d). Por isso o estranhamento quando o clítico de primeiro grupo se posiciona após o verbo (95b) e o do segundo grupo se posiciona antes do verbo (95c):

(95)
a. A Maria fez isso só para **me** magoar.
b. ? A Maria fez isso só para magoar-**me**.

c. ?? Com o intuito de **a** agradar, o João mandou flores.
d. Com o intuito de agradá-**la**, o João mandou flores.

(Kanthack, 2002: 116-117)

O mesmo pode ser atestado em dados do *corpus* consultado: os clíticos de primeiro grupo *lhe* (96a) e *nos* (96b) se encontram antes do verbo não-finito e os clíticos de segundo grupo *o* (97a) e *a* (97b) se encontram na posição pós-verbal.

(96)
a. ...tenderá a **lhe** *dar* apoio...
b. ...só temos razões para confiar e **nos** *unirmos* sempre mais.

(Corpus: CETENFolha)

(97)
a. Já o desafiei, pelo rádio, a fornecer nomes, para *ajudá-**lo*** a combater...
b. ...ajudar uma pessoa a relaxar é *ensiná-**la*** a reconhecer e se apoiar...

(Corpus: CETENFolha)

Essa característica é tão forte que mesmo em frases com palavras negativas antes do verbo, os clíticos do primeiro grupo são preferencialmente pré-verbais (98a), mas os do segundo grupo continuam a ser mais naturais na posição pós-verbal (98d), como é possível verificar ao se comparar (98c) e (98d):

(98)
a. Para *não* **me** cansar, fui pelo caminho mais curto.
b. * Para *não* cansar-**me**, fui pelo caminho mais curto.
c. ?? Para *não* **o** assustar, os soldados se retiraram.
d. Para *não* assustá-**lo**, os soldados se retiraram.

(Kanthack, 2002: 117)

Essas ideias de Kanthack (2002) acerca da posição dos clíticos em frases com palavras negativas também foram observadas no *corpus* consultado. Em (99a) e (99b), os clíticos de primeiro grupo são proclíticos, ocorrendo entre o advérbio negativo e o verbo finito. Já o clítico de segundo grupo é enclítico em (99c) e (99d), apesar da presença da palavra *não* antes do verbo finito.

(99)
a. *Não* **vos** peço clemência, *não* **vos** peço espírito...
b. Essa análise *não* **me** cabe.
c. ...tinha lugares vazios (por que *não* ocupá-**los** com...
d. ... nunca fez nada ruim, não há como *não* cantá-**lo** direito.

(Corpus: CETENFolha)

No caso das sentenças com dois ou mais verbos adjacentes, os clíticos do primeiro grupo se posicionam antes do verbo mais baixo (principal), como em (100a), caso contrário pode causar estranhamento, como em (100b). Os excertos seguintes exemplificam orações com verbos no infinitivo:

(100)
a. Ela quer **me** encontrar nas férias.
b. ? Ela quer encontrar-**me** nas férias.

(Kanthack, 2002: 119)

Segundo a autora, "o clítico deve vir à esquerda do último verbo do conjunto, que pode estar no infinitivo, no gerúndio ou no particípio" (Kanthack, 2002: 122). Essa afirmação pode ser verificada nos excertos a seguir em que foram apresentadas construções com verbos no gerúndio (101) e no particípio (102). As estruturas mais aceitáveis são as que apresentam o clítico de primeiro grupo na posição proclítica ao verbo principal, como em (101a) para o gerúndio e (102a) para o particípio:

(101)
a. Ele está **me** enrolando há vários dias.
b. ?Ele está enrolando-**me** há vários dias.
c. ??Ele **me** está enrolando há vários dias.
d. *Ele está-**me** enrolando há vários dias.

(Kanthack, 2002: 121)

(102)
a. Ela já tinha **me** procurado outras vezes.
b. *Ela já tinha procurado-**me** outras vezes.
c. ??Ela já **me** tinha procurado outras vezes.
d. *Ela já tinha-**me** procurado outras vezes.

(Kanthack, 2002: 122)

Esse comportamento dos clíticos de primeiro grupo em sequências verbais também pode ser verificado em sentenças do *corpus* consultado com verbos no infinitivo (103a), gerúndio (103b) e particípio (103c):

(103)
a. ...a sua companhia *quer **me** vender*...
b. ...a borrasca, *fomos **nos** aproximando* das novas vertentes...
c. ...eles *terem **me** abandonado* em 1940...

(Corpus: CETENFolha)

Já os clíticos do segundo grupo não podem estar antes do verbo não-finito, hipótese que é confirmada ao se observar a agramaticalidade dos seguintes com verbos no infinitivo (104a), no gerúndio (104b) e no particípio (104c):

(104)

a. *A Maria pode **o** encontrar naquele bar amanhã.
b. *A Maria está **o** encontrando naquele bar nesse momento.
c. *A Maria tinha **o** encontrado naquele bar ontem.

(Kanthack, 2002: 124)

Contudo, esses clíticos podem ocorrer antes do verbo finito, mas causando grande estranhamento como em (105):

(105)

a. ??A Maria **o** pode encontrar amanhã.
b. ??A Maria **o** está esperando nesse momento.
c. ??A Maria **o** tinha encontrado ontem.

(Kanthack, 2002: 124)

Kanthack (2002: 124) afirma que esse tipo de clítico ocorre preferencialmente em ênclise ao verbo principal não-finito, como é possível verificar em (106b) e (106d):

(106)

a. *A Maria pode querê-**lo** encontrar.
b. A Maria pode querer encontrá-**lo**.
c. *A Maria vai podê-**lo** encontrar.
d. A Maria vai poder encontrá-**lo**.

(Kanthack, 2002: 125)

Essa autora afirma que "em sentenças com dois ou mais verbos adjacentes vimos que o clítico *o* pode se posicionar em dois lugares" (Kanthack, 2002: 125), ou seja, em ênclise ao verbo principal não-finito ou em próclise ao verbo auxiliar finito. Entretanto, ao comparar os exemplos (105), em que há grande estranhamento com o clítico antes do verbo finito; e os exemplos (106), em que se observa a maior naturalidade de frases com o clítico após o verbo principal não-finito, a autora constata que "o melhor lugar para ele ocorrer é depois do verbo infinitivo e gerúndio" (Kanthack, 2002: 125).

Também foram encontrados no *corpus* consultado exemplos de sequência verbais com o verbo principal infinitivo e o clítico de segundo grupo após o verbo mais baixo, como em *comprometê-los* (107a) e *impedi-lo* (107b) e *buscá-los* (107c).

(107)

a. ...conseguido via FSE, não *quer comprometê-**lo*** através de rombos...
b. O que a MP *poderia impedi-**lo*** de fazer, nesse caso...
c. ...se livrar da prisão, *vai mandar buscá-**los*** imediatamente para Maceió...

(Corpus: CETENFolha)

No caso do particípio, o contraste entre os exemplos (110a) e (110b) apresentados a seguir demonstra que "o único caso em que *o* deve ser licenciado numa posição mais alta (embora seja uma sentença marginal) é aquele em que está presente o particípio" (Kanthack, 2002: 125). Por um lado, a posição enclítica é a preferencial para o clítico de segundo grupo em frase com verbo no infinitivo (108a) e no gerúndio (109a). Por outro lado, com o verbo no particípio a posição enclítica é agramatical, sendo preferível inserir o verbo antes do verbo finito (110b).

(108) Verbo no infinitivo

a. A Maria *pode encontrá-***lo** amanhã.
b. ??A Maria **o** *pode encontrar* amanhã.

(109) Verbo no gerúndio

a. ?A Maria *está esperando-***o** nesse momento.
b. ??A Maria **o** *está esperando* nesse momento.

(110) Verbo no particípio

a. *A Maria *tinha encontrado-***o** ontem.
b. ??A Maria **o** *tinha encontrado* ontem.

(Kanthack, 2002: 125)

Kanthack resume a questão da seguinte maneira: "vimos que em sentenças com dois ou mais verbos adjacentes a colocação do clítico *o* não se pauta pela dos demais. Quando é usado, o clítico *o* poderá ocorrer em duas posições: antes do verbo finito ou depois do verbo não-finito, exceto com o particípio ativo" (Kanthack, 2002: 125).

No caso das sentenças de particípio passivo, Kanthack (2002) ressalta que "a cliticização somente acontece em uma posição mais alta (...) o clítico se posiciona junto a um verbo finito, e não junto de um verbo de natureza não-finita" (Kanthack, 2002: 127). Diante disso, como é possível observar nos excertos em (111), a melhor posição para o clítico "*me*" é a proclítica ao verbo finito (111c), independentemente de o verbo "ser" no infinitivo ser substituído por um verbo no particípio ou gerúndio.

(111)

a. * Esta casa vai ser **me** dada de presente.
b. * Esta casa vai **me** ser dada de presente.

c. ?? Esta casa **me** vai ser dada de presente.

(Kanthack, 2002: 128)

Tendo em vista esse tipo de construção ser inacusativa, "os clíticos acusativos não são licenciados (...) os clíticos de segundo grupo jamais poderão ocorrer, já que eles são de natureza acusativa" (Kanthack, 2002: 129).

O comportamento distinto dos clíticos de primeiro e segundo grupo é evidenciado quando Kanthack (2002: 129) analisa os clíticos em início de sentença. Ao comparar os excertos (112a) e (112c), é possível verificar que os clíticos do primeiro grupo podem surgir na posição inicial da frase, mesmo na ausência de elementos fonéticos à sua esquerda, mas os do segundo grupo não podem.

(112)

a. **Te** procuro todos os dias.
b. ? Procuro-**te** todos os dias.
c. * **O** procuro todos os dias.
d. ? Procuro-**o** todos os dias.

(Kanthack, 2002: 129)

O clítico "*o*" somente pode ocorrer antes do verbo quando há elementos foneticamente realizados, como em (113a). Nesse caso, o comportamento dos clíticos desse grupo é semelhante aos do primeiro grupo (113c). Segundo Kanthack "esses exemplos mostraram, portanto, é que a lei Tobler-Mussafia favorece o uso da ênclise apenas quando o *o* for o clítico em questão" (Kanthack, 2002: 130)[18].

(113)

a. O seu pai **o** procura todos os dias.
b. ? O seu pai procura-**o** todos os dias.
c. O seu pai **te** procura todos os dias.
d. ? O seu pai procura-**te** todos os dias.

(Kanthack, 2002: 130)

Também no *corpus* consultado é possível encontrar os clíticos de primeiro grupo na posição inicial da frase, como exemplificam os clíticos "*te*" (114a) e "*me*" (114b) a seguir:

[18] A lei Tobler-Mussafia "determina que em línguas românicas antigas os clíticos não podem ocorrer na posição inicial da sentença. Se nenhum constituinte aparece antes do verbo, o clítico deve ser licenciado na posição pós-verbal" (Kanthack, 2002: 23). Esse fenômeno ainda ocorre em PE, entretanto, em algumas línguas românicas modernas, entre elas o PB, o clítico pode ocorrer em posição inicial. Para mais informações acerca desse fenômeno, também cf. Duarte (2003: 849).

(114)

a. **Te** chamaram para sempre de ex-presidiário.
b. **Me** contaram o motivo e eu não acreditei...

(Corpus: CETENFolha)

Além de comprovarem que a próclise é a ordem mais comum em PB para a maioria dos clíticos, esses estudos demonstram que o comportamento desses pronomes pode mudar consoante a forma dos clíticos, o tipo de verbo e a sua inserção em sequências verbais. Como ressalta Duarte (2015), é necessário lembrar que em diversos textos escritos em PB há um uso exagerado de ênclise mesmo em contextos em que a próclise é considerada obrigatória por gramáticas normativas. Todos esses aspectos serão considerados durante a análise dos erros envolvendo clíticos para tentar compreender as escolhas feitas pelos editores e anotadores durante o processo de pós-edição e anotação dos dados analisados.

5.2.1.3 Ordem dos advérbios

Os elementos inseridos na classe dos advérbios podem ter escopo sobre diversos tipos de constituintes, incluindo sentenças ou sintagmas, auxiliando na composição semântica da frase na qual estão inseridos. Por exemplo, em (115a) o advérbio "*não*" modifica o predicado, contribuindo no estabelecimento do valor de verdade acerca do que foi dito. Já em (115b), os advérbios "*ontem*" e "*anteontem*" auxiliam a situar temporalmente a situação descrita.

(115)

a. Isso **não** vem à toa.
b. ...marcou 16 pontos **anteontem** e outros 25 **ontem**.

(Corpus: CETENFolha)

Os advérbios podem aparecer nas sentenças desacompanhados, como nos exemplos em (115), ou acompanhados por especificador e/ou complementos. Nesse último caso, eles se inserem num sintagma adverbial (SAdv) e, como os outros tipos de sintagma, seguem a estrutura Especificador + Núcleo + Complemento. Enquanto seu núcleo é sempre adverbial, o seu especificador é geralmente um advérbio e o seu complemento pode ser um sintagma preposicional (SP) que pode introduzir um SN ou uma oração (Raposo, 2013: 1583). As sentenças apresentadas a seguir podem exemplificar casos de SAdv especificados por outro advérbio (116a) ou complementadas por SP (116b).

(116)

a. Barros morreu *mais **tarde*** em acidente aéreo.
b. «Voto nas pessoas, ***independentemente** de partidos*».

(Corpus: CETENFolha)

De maneira geral, os advérbios podem ocorrer em variadas posições na sentença e, por isso, apresentam uma distribuição bastante livre. Para exemplificar, o advérbio "*ontem*" assinalado nos exemplos (117) a seguir pode ocupar diversas posições da sentença.

(117)

a. **Ontem**, o Pedro tinha lido o livro à avó.
b. O Pedro, **ontem**, tinha lido o livro à avó.
c. O Pedro tinha, **ontem**, lido o livro à avó.
d. O Pedro tinha lido, **ontem**, o livro à avó.
e. O Pedro tinha lido o livro, **ontem**, à avó.
f. O Pedro tinha lido o livro à avó **ontem**.

(Costa, 2008: 77)

Como ressalta Costa (2008: 78), a ordem dos advérbios está intimamente ligada à função e ao sentido desses advérbios na frase. Isso pode ser verificado ao comparar os exemplos (118) e (119): a frase com o advérbio "*completamente*", modificador do predicado SV, é agramatical quando esse termo está no início da frase (118a) ou entre o sujeito e o predicado (118b), mas essas mesmas posições estão disponíveis para o advérbio "*felizmente*", que modifica a frase, em (119a) e (119b). Note-se que as frases (119c) e (119d) são possíveis, mas, nesse caso, o advérbio veicula informação distinta, já que é orientado para a maneira como o João leu o livro (orientado para o agente) e não é modificador de frase.

(118)

a. ***Completamente**, o João leu o livro.
b. *O João, **completamente**, leu o livro.
c. O João leu **completamente** o livro.
d. O João leu o livro **completamente**.

(119)

a. **Felizmente**, o João leu o livro.
b. O João, **felizmente**, leu o livro.
c. *O João leu **felizmente** o livro.
d. *O João leu o livro **felizmente**.

(Costa, 2008: 78)

Por conseguinte, procurou-se nos próximos parágrafos fazer uma descrição dos advérbios que aborde brevemente as funções que eles podem exercer na frase e a classificação semântica desses elementos com o objetivo de encontrar uma descrição geral da maneira como os advérbios se ordenam nas frases. Serão apresentados dados do *corpus*, no sentido de mostrar que as descrições para o PE são adequadas ao PB.

5.2.1.3.1 Funções dos advérbios

Como veremos nos parágrafos seguintes, os advérbios podem exercer diferentes funções na sentença, sendo a função de adjunto a mais típica dessa classe gramatical. Quando são adjuntos, os advérbios veiculam informação opcional para a sentença, adicionando informação adicional à frase, por isso podem ser omitidos sem prejudicar a coerência (Raposo, 2013: 1570). O advérbio assinalado em (120a) é um exemplo de adjunto que, apesar de adicionar informação semântica à frase, é opcional e pode ser omitido sem prejudicar a aceitabilidade da frase (120b).

(120)

a. De braços cruzados, ela acompanhava **atentamente** o marido.
b. De braços cruzados, ela acompanhava o marido.

Os advérbios, geralmente os que denotam quantidade ou de grau, também podem especificar adjetivos (121a), advérbios (121b) ou verbos. Ao exercer essa função os advérbios se posicionam geralmente à esquerda de seu núcleo, como em "muito tranquilamente" e "bastante claro".

(121)

a. ...com o acesso às informações ***bastante*** *claro*.
b. ...assumo ***muito*** *tranquilamente* o meu histórico político.

(Corpus: CETENFolha)

Raposo (2013: 1594) aponta que alguns advérbios podem ser selecionados por um predicador verbal para completar o valor semântico da frase. Nos casos em que o advérbio selecionado representa uma entidade que participa da situação descrita, ele é considerado um dos argumentos do predicador e, como outros complementos, não pode ser omitido da frase. Por exemplo, considerando o verbo "vive" com o sentido de morar, em (122b) a omissão do advérbio "*aqui*" faz com que a frase seja agramatical.

(122)

a. O fundador do restaurante já não *vive* ***aqui***.
b. *O fundador do restaurante já não *vive*.

(Corpus: CETENFolha, com modificações em (122b))

Nos casos em que os advérbios selecionados não representam participantes da situação descrita, eles são denominados "quase argumentos". Assim como os argumentos do predicado, eles não podem ser omitidos pois são elementos centrais na leitura semântica da frase. Veja-se o exemplo (123b), em que a omissão do advérbio "*bem*" com função de "quase argumento" leva à agramaticalidade da frase.

(123)

a. Só disse que o treino *correu* ***bem***.
b. *Só disse que o treino *correu*.

(Corpus: CETENFolha, com modificações em (123b))

Os advérbios também podem ser predicadores em orações copulativas. Nesses casos, eles são os elementos centrais da frase e impõem restrições semânticas sobre o sujeito. Os diferentes advérbios assinalados nas orações copulativas em (124) são exemplos de predicadores.

(124)

a. ...o grão *parece* ***bem***.
b. A notícia que ele procurava *estava* ***lá***.
c. ...o desempate *é* ***hoje***.

(Corpus: CETENFolha)

5.2.1.3.2 Classificação semântica dos advérbios

Quanto à classificação semântica dessa classe gramatical, Costa (2008) distingue entre os advérbios modificadores de predicado e os modificadores de frase. Enquanto os advérbios do primeiro grupo têm escopo sobre o SV, os do segundo grupo modificam a frase inteira. O advérbio assinalado em (125a) modifica o SV enquanto o advérbio "*infelizmente*" em (125b) modifica toda a frase.

(125)

a. Seus trabalhadores, mais educados, *apreendem* ***rapidamente*** *as novas tecnologias.*
b. Na Argentina, **infelizmente**, o anti-semitismo é muito maior que no sul do Brasil.

(Corpus: CETENFolha)

Costa (2008) divide os modificadores de predicado em três subclasses: os advérbios de localização, que podem situar o predicado no tempo (126a) ou no espaço (126b); os advérbios de modo (126c), que fornecem informação sobre a maneira como o estado de coisas descrito no SV se desenvolve; e os advérbios de quantidade e grau que denotam uma quantificação ou a intensidade (126d) com a qual se manifesta o predicado (Costa, 2008: 43).

(126)

a. ...promove **amanhã** em seu auditório...
b. «Não aguenta mais viver **longe da família** e deve retornar».
c. ...foi confeccionado **manualmente**, tem 20 páginas...
d. ...com a raça cresceram **bastante** nos últimos anos.

(Corpus: CETENFolha)

Os modificadores de frase, por sua vez, são divididos em dois grupos: os advérbios avaliativos, que exprimem uma avaliação sobre o conteúdo da proposição (127a); e os advérbios conectivos (127b), que estabelecem relações entre frases ou constituintes, funcionando textualmente (Costa, 2008: 53-62).

(127)

a. No caso de Lula, **francamente**, o que houve foi inabilidade...
b. ...servem para o cálculo do custo e, **consequentemente**, para a fixação...

(Corpus: CETENFolha)

Há advérbios que não modificam necessariamente a frase ou o predicado, mas se associam a outros constituintes. Estão inseridos nesse conjunto os advérbios de quantidade e grau, quando esses modificam outros constituintes que não são o predicado. Em (128a), por exemplo, o advérbio de intensidade especifica o adjetivo "intenso". Os advérbios focalizadores também podem realçar outros constituintes da sentença além do predicado. Isso pode ser observado em (128b), onde o advérbio focalizador tem escopo sobre "*Natal e Mossoró*". Os advérbios polarizadores[19] "não" e "sim", além de marcarem o valor de verdade da frase, podem modificar constituintes não frásicos como em (128c) e (128d). (Costa, 2008: 63-73).

[19] Considerando-se os dados a serem analisados, não serão apresentadas explicações acerca da ordem dos advérbios polarizadores. Para mais detalhes, cf. Costa (2008).

(128)

a. ...não ocorria resfriamento ***tão*** *intenso*.
b. Atualmente, ***só*** *Natal e Mossoró* possuem...
c. ....um fundo de investimento que teria, *ele* ***sim***, a possibilidade...
d. ...para alguns, mas ***não*** *desenvolvimento social*», afirmou...

(Corpus: CETENFolha)

5.2.1.3.3 Ordem dos advérbios segundo sua classificação semântica

A associação entre os advérbios e os outros constituintes da frase pode mudar consoante a posição desses elementos. Segundo Costa (2008), de modo geral, os advérbios são pré-verbais quando modificam a frase e pós-verbais quando se associam ao predicado.

Os modificadores de frase ocorrem tipicamente em posição pré-verbal, por isso podem ser inseridos no início da frase com facilidade. No caso da posição entre o sujeito e o predicado, Costa (2008) afirma ser necessária uma entoação parentética[20], muitas vezes marcada por vírgulas na escrita. O advérbio avaliativo "*infelizmente*" e o advérbio conectivo "*contudo*" apresentados nas frases a seguir exemplificam essas duas posições: no início da frase, em (129); e entre o sujeito e o predicado (130).

(129)

a. **Infelizmente** alguns direitos constitucionais serão prejudicados.
b. **Contudo**, o deputado teve que passar pelo interrogatório direto e oral.

(130)

a. O custo de fazer a transição política antes da econômica **infelizmente** existe.
b. Seus assessores, **contudo**, repetiram a explicação dada por Dias.

(Corpus: CETENFolha)

Para que os modificadores de frase estejam em posição pós-verbal, Costa (2008) ressalta que é necessário haver uma curva melódica particular. Os advérbios avaliativos e conectivos devem ter uma entoação parentética quando estão imediatamente após o verbo (131) e uma pausa, tipicamente representada por vírgulas na escrita, quando estão na última posição da frase (132).

[20] As construções parentéticas são expressões que apesar de estarem linearmente presentes num enunciado e, em termos de conteúdo, com ele direta e indiretamente relacionadas, aparentam ser estrutural e prosodicamente independentes. (Cunha e Moita, 2011)

(131)

a. Reagiu com paixão, como se diz **certamente** por aí.
b. Não haveria **portanto** pressão no orçamento.

(132)

a. Não se discutiu filigranas jurídicas, **certamente**.
b. Na pesquisa anterior, esse número era de 44 %, praticamente sem variação, **portanto**.

(Corpus: CETENFolha)

A posição típica dos modificadores de predicado é após o verbo, como exemplificam os advérbios de modo, localização temporal e espacial assinalados em (133a), (133b) e (133c), respectivamente:

(133)

a. Mas o preconceito se manifestou **depressa**.
b. Duas informações sobre a agricultura circularam **recentemente**, causando efeitos...
c. Ele almoçou **ali**, após se reunir...

(Corpus: CETENFolha)

Os advérbios de quantidade e grau também ocorrem tipicamente após o verbo quando são modificadores do SV. Contudo, esse tipo de advérbio não se posiciona facilmente após complementos (Costa, 2008: 81). Isso pode ser verificado ao mudar a posição do advérbio "*bastante*": a sentença em (134b) causa estranhamento devido à distância entre o verbo e o advérbio.

(134)

a. ....a baixa afluência de público *dificultou **bastante** a compra de animais*.
b. ? ...a baixa afluência de público *dificultou a compra de animais **bastante***.

(Corpus: CETENFolha, com modificações em (134b))

Segundo Costa (2008: 90-91), os advérbios homônimos de adjetivos são modificadores de predicado, por isso podem estar em posição pós-verbal sem pausa, como o advérbio "*alto*" em (135a). Ainda assim, não ocorrem com facilidade na última posição da frase após complementos, sendo mais adequada a posição logo após o verbo. Isso pode ser verificado no exemplo (135b), em que, ao mudar a posição do advérbio "*alto*", a frase passa a causar certo estranhamento.

(135)

a. Os pefelistas negam, mas *cobrarão **alto** a fatura*.
b. ?? Os pefelistas negam, mas *cobrarão a fatura **alto***.

(Corpus: CETENFolha, com modificações em (135b))

Tendo em vista que os modificadores de SV tipicamente não ocorrem em posição pré-verbal, a maior parte dos advérbios desse grupo só ocorre na posição inicial da frase em contextos contrastivos, exercendo função de tópico da frase (Costa, 2008: 82). Os advérbios de modo (136a), de localização espacial (136b) e de quantidade e grau (136c) assinalados respectivamente nas sentenças seguintes exemplificam contextos em que isso ocorre:

(136)

a. **Calmamente** Suplicy foi cumprimentando os presentes...
b. **Lá**, 98,56 % dos animais foram vacinados.
c. **Pouco** viajou naquela campanha, não preparou material de propaganda...

(Corpus: CETENFolha)

Costa (2008: 83) aponta dois outros casos em que a posição inicial é possível: em contextos apresentativos com advérbios de localização espacial e inversão sujeito-verbo (137a) e em contextos exclamativos com inversão sujeito-verbo (137b).

(137)

a. **Aqui** reside o nó da questão: (...)

(Corpus: CETENFolha)

b. **Tanto** comi eu naquela festa!

(Costa, 2008: 83)

Os advérbios de localização temporal apresentam um comportamento distinto, pois podem ocorrer nessa posição sem necessidade que haja necessariamente um contraste. No exemplo (138a), há uma pausa, marcada pela vírgula entre o advérbio e o resto da frase, enquanto no exemplo (138b) não foi inserida vírgula.

(138)

a. «**Antigamente**, havia barreiras de fiscalização do Estado...
b. **Hoje** instituições respeitadas no mundo inteiro atestam a teoria...

(Corpus: CETENFolha)

A posição entre o sujeito e o predicado está disponível para a maioria os advérbios de modo e os de quantidade e grau, mas nessa posição têm tendência a apresentarem uma

leitura de advérbio avaliativo. Isso pode ser verificado nos seguintes exemplos com os advérbios de modo (139a) e de intensidade (139b) inseridos antes do SV:

(139)

a. ?«O presidente ***perfeitamente*** *sabe que tudo é feito com fraternidade e clareza*».
b. ? Apesar de contar com uma oferta de boa qualidade, a baixa afluência de público ***bastante*** *dificultou a compra de animais*.

(Corpus: CETENFolha, com adaptações)

Os advérbios de localização espacial com localização definida (140a) e os advérbios de localização temporal (140b) podem ocorrer entre o sujeito e o predicado, desde que exista entoação parentética (Costa, 2008: 85).

(140)

a. Um advogado, um padre, um juiz, um promotor, tudo ***ali*** *tem seu preço*.
b. «E se o funcionário ***frequentemente*** *ultrapassar suas metas*, terá direito a uma promoção."

(Corpus: CETENFolha)

Os valores semânticos dos advérbios podem mudar dependendo do contexto ou da posição desses elementos na sentença. Essa característica pode ser comprovada ao interpretar os exemplos citados a seguir. Em (141a) o termo "inteligentemente" é interpretado como um advérbio modificador de frase, mas em (141b) o mesmo termo modifica o predicado.

(141)

a. Inteligentemente, a polícia interrogou o João
   (= Foi inteligente da parte da polícia interrogar o João)
b. A polícia interrogou inteligentemente o João.
   (= A polícia interrogou de maneira inteligente o João)

(Costa, 2008: 78-79)

Além disso, a ordem de alguns advérbios pode causar ambiguidade no sentido da sentença, sendo necessário mais contexto para compreender o real sentido da frase. Essa possível ambiguidade devido à posição pode ser observada em frases como a do exemplo (141), que possui mais de uma possível interpretação. Outras vezes, a posição dos advérbios na frase é gramatical, mas pode causar certo estranhamento no leitor, como foi observado em diversos exemplos já apresentados nos parágrafos anteriores. Todos

esses aspectos são importantes para o processo de tradução e produção textual, pois estão intimamente ligados à coesão e coerência textual. Por isso, procurou-se considerar esses diferentes pontos durante a análise dos dados anotados.

Quando não modificam o predicado, os advérbios focalizadores[21] se inserem tipicamente imediatamente à esquerda do elemento que modificam. Para exemplificar, o advérbio "*só*" modifica o possessivo "*nosso*" em (142a). Na posição final da sentença, os focalizadores têm escopo sobre o predicado se houver uma pausa (Costa, 2008: 94), como se pode observar em (142b).

(142)
a. ...fazemos um desenho ***só*** *nosso*? "
b. Há problemas de didatismo, **apenas**.

(Corpus: CETENFolha)

É de notar que os focalizadores inseridos logo antes do verbo modificam todo o SV ou o último constituinte da frase. Por vezes, é necessário mais contexto para saber qual desses dois constituintes são modificados pelo advérbio. Para exemplificar, em (143), há duas interpretações possíveis: o advérbio "apenas" pode modificar o SV inteiro (143a) ou somente o SN "os pulsos da ex-mulher" (143b).

(143) ...não houve agressão de fato e ele ***apenas*** *segurou [os pulsos da ex-mulher]*. (Corpus: CETENFolha)

a. (= Segurar os pulsos da ex-mulher foi a única coisa que ele fez)
b. (= Ele apenas segurou os pulsos da ex-mulher e não segurou outra coisa)

5.2.1.3.4 Ordem dos advérbios em sequências verbais

Além das posições apresentadas, os advérbios podem se posicionar entre verbos que se encontram em sequência. Costa (2008: 101) ressalta a importância de observar não somente o tipo de advérbio, mas também o formato da sequência para verificar se é possível inserir um advérbio entre os verbos.

No caso de sequências com verbos auxiliares e modais, de maneira geral, os modificadores de predicado podem ocorrer entre o verbo auxiliar e o verbo principal com formatos "*ter* + *particípio passado*", "*ir* + *infinitivo*", e "*haver de* + *infinitivo*" (Costa, 2008: 101).

[21] Para mais detalhes acerca da posição advérbios focalizadores, cf. os trabalhos de Costa (2008).

(144)

a. O coronel *tinha **já** acabado* seu exercício quando foi atacado.
(Corpus: CETENFolha, com modificações nossas)
b. O coronel *vai **amanhã** acabar* seu exercício.
c. O coronel *há de **aqui** fazer* seu exercício.

Os advérbios de modo e intensidade, entretanto, ocorrem preferencialmente após o particípio e o infinito. Por isso os advérbios de modo e intensidade nos exemplos (145b) e (146b), respectivamente, causam certo estranhamento quando se encontram após o particípio e o infinitivo.

(145)

a. Espero que os Frias não o *tenham lido **rapidamente**.*
b. ? Espero que os Frias não o *tenham **rapidamente** lido.*

(146)

a. ? O Tribunal *vai apreciar **muito*** sua competência..
b. O Tribunal *vai **muito** apreciar* sua competência..
(Corpus: CETENFolha, com modificações em (145b) e (146a))

O mesmo comportamento dos advérbios modificadores de SV é observado em sequências de dois verbos, sendo o primeiro um verbo modal. Por um lado, os advérbios de tempo e espaço podem facilmente ser inseridos entre os dois verbos:

(147)

a. ...coordenador *pode **hoje** ser* o primeiro homem...
b. ...quanto ao que *pode **aqui** ocorrer* de positivo...
(Corpus: CETENFolha)

Por outro lado, os advérbios de modo e intensidade entre os elementos da sequência verbal podem causar estranhamento. Verifica-se isso ao comparar as frases (148a) e (148b), em que a inserção do advérbio de intensidade após o segundo verbo parece a ser a ordem mais natural em português.

(148)

a. ...apenas os candidatos a presidente e a vice *podem aparecer **muito***.
b. * ...apenas os candidatos a presidente e a vice *podem **muito** aparecer*.
(Corpus: CETENFolha, com alterações em (148b))

Nas frases passivas, com o formato "verbo auxiliar *ser* + particípio", os advérbios modificadores de predicado, incluindo os de modo e intensidade que não são advérbios

derivados de adjetivos, podem preceder ou seguir o particípio. As frases em (149) exemplificam advérbios de localização temporal (149a) e de modo (149b) inseridos entre os verbos da voz passiva. Ao confrontar as frases em (150), verifica-se que a inserção do advérbio homônimo de adjetivo "*rápido*" causa a agramaticalidade da sentença em (150b).

(149)

a. Numa das maiores transações conduzidas na Argentina nos últimos anos, *foi* ***ontem*** *anunciado*, em Buenos Aires...
b. O tema do aborto *foi* ***moderadamente*** *incluído* no programa de governo do PT...

(150)

a. ...o projeto ideal para *ser* ***rapidamente*** *aprovado* pelo Congresso...
b. * ...o projeto ideal para *ser* ***rápido*** *aprovado* pelo Congresso...

(Corpus NILC/São Carlos: CETENFolha, com alterações em (150b))

Os advérbios modificadores de frase geralmente podem ocorrer entre dois verbos desde que exista entoação parentética. Os advérbios avaliativo (151a) e conectivo (151b) nas frases a seguir exemplificam essa possibilidade:

(151)

a. A lista detalhada *será* ***provavelmente*** *divulgada* hoje.
b. A lista detalhada *será,* ***contudo****, divulgada* hoje.

(Corpus: CETENFolha)

Os advérbios focalizadores também podem ocorrer entre dois verbos sem necessitar de entoação parentética como se pode observar em (152a). Segundo Costa (2008: 104), nesse caso se considera que modificam o verbo à direita.

(152)

a. ...Pagliuca *tinha* ***só*** *agarrado* um tiro de longe...

(Corpus: CETENFolha)

No caso de sequências com dois verbos principais, segundo Costa (2008: 104), qualquer classe de advérbio pode se posicionar entre os dois verbos, inclusive os advérbios de modo (153a), intensidade (153b) e conectivo (153c), como exemplificam os seguintes excertos:

(153)

a. ...Autolatina, que já *decidiu* ***oficialmente*** *urvizar* os preços de toda sua linha de veículos...
b. ...na Europa e *queria* ***muito*** *jogar* por um time...
c. O deputado (...) *assegura* ***também*** *ser* contrário...

(Corpus: CETENFolha)

Quando ocorrem entre dois verbos principais, os modificadores de predicado necessariamente se associam ao primeiro verbo. Na sentença (154), por exemplo, o advérbio de modo modifica o verbo à sua esquerda, por isso a interpretação mais adequada de (154) é a que se apresenta em (154a).

(154)

...Autolatina, que já *decidiu* ***oficialmente*** *urvizar* os preços de toda sua linha de veículos...

(Corpus: CETENFolha)

a. (= A decisão foi feita de maneira oficial)
b. ($\neq$ A urvização foi feita de forma oficial)

Já os advérbios focalizadores podem se ligar a ambos os verbos principais, por isso muitas vezes o sentido da frase pode ser ambíguo. Em (155), por exemplo, duas interpretações são possíveis: em (155a) o advérbio "*apenas*" se liga ao verbo à sua esquerda e em (155b) ele focaliza o verbo "*mostrar*".

(155)

A Coréia do Sul *quer* ***apenas*** *mostrar* um bom futebol para aumentar as suas chances...

(Corpus: CETENFolha)

a. (= Mostrar um bom futebol é a única coisa que a Coreia do Sul quer fazer)
b. (= A Coreia do Sul vai apenas mostrar um bom futebol e não fazer outra coisa para aumentar suas chances)

5.2.2 Ordem de palavras internamente ao SN

Como se afirmou na secção 5.1.1, o núcleo do SN, o nome, é a parte central e fundamental desse sintagma. Sendo assim a colocação dos outros elementos se organiza de acordo com as suas propriedades. Segundo Raposo e Miguel (2013: 710), os nomes podem se combinar com complementos (156a), modificadores (156b) e especificadores (156c). Esses elementos se organizam dentro do SN numa ordem particular, sendo a ordem básica [Especificador] + [Núcleo] + [Modificador] / [Complemento].

(156)

a. ...a aplicação **de indexador**...
b. ...chuva **forte**...
c. ...**muitos** filhotes...

(Corpus: CETENFolha)

Os parágrafos seguintes se dedicam à descrição da ordem dos especificadores, modificadores e complementos em relação ao núcleo do SN. Tendo em vista o comportamento particular de certos SAs, será dada atenção especial a esse tipo de modificador.

5.2.2.1 Ordem dos especificadores

De acordo com Raposo e Miguel (2013: 723), quando o SN contém somente um especificador[22], este aparece na posição inicial do SN, à esquerda do nome, como nos exemplos (157). Nesta posição pré-nominal, podem ocorrer quantificadores vagos e numerais com uma interpretação indefinida.

(157)

a. ...fez com que (**muitos**/**poucos**/**vários**/**bastante**/**cinco**) *contribuintes* só pagassem...
b. (**Os**/**Uns**/**Estes**/**Alguns**) *contribuintes* reclamam...

(Corpus: CETENFolha, com modificações nossas)

Acerca da ordem dos possessivos no SN, Raposo e Miguel (2013) afirmam que os mesmos se posicionam à esquerda do nome quando o SN é definido (158a) e à direita se o SN é indefinido (158b). Os possessivos também podem preceder um quantificador ou um numeral, como em (158c).

(158)

a. ...*a* **sua** participação no Campeonato...
b. *Um* aluno **meu** do curso...
c. *A* **minha** *pouca* experiência foi compensada...

(Corpus: CETENFolha)

Outra particularidade é a aceitação, por parte de alguns falantes, da ocorrência de *um* e *algum* com possessivo em posição pré-nominal, como é possível verificar em (159).

[22] Acerca da ordem de SN com mais de um especificador, cf. Raposo e Miguel (2013).

Segundo Raposo e Miguel (2013), essa característica aproxima o artigo indefinido e o determinante *algum* do artigo definido e dos demonstrativos.

(159)
a. ...recuperou **um** *seu* antigo caso...
b. E também **alguns** *seus* master-franqueados...

(Corpus: CETENFolha)

Segundo Raposo e Miguel (2013), um possessivo pré-nominal não ocorre na posição inicial do SN em PE, sendo sempre acompanhado por um determinante definido, como em (158a). No entanto, a mesma regra não se aplica ao PB, em que o pronome possessivo pode ocorrer na posição inicial do SN, sem qualquer determinante, como é possível observar nos excertos assinalados em (160), retirados do *corpus* consultado.

(160)
a. ...ainda compram **nossos** *touros* para produção...
b. ...ano levei **minha** *filha de 5 anos* para ser consultada...
c. ...tentou recompor **seu** *time*.

(Corpus: CETENFolha)

5.2.2.2 Ordem dos modificadores

A função dos modificadores do nome é, como já se referiu em 5.1.1.2, introduzir propriedades adicionais na denotação do nome, ou da expressão nominal constituída pelo nome e seu(s) complemento(s). Assim como os complementos, os modificadores se manifestam tipicamente em posição pós-nominal (à direita). Apresentamos abaixo exemplos de algumas das possibilidades de modificadores de nome: SA (161a), SP (161b) e orações, neste caso uma oração relativa (161c).

(161)
a. um gato **siamês**
b. o gato **da minha vizinha**
c. o jogador **que marcou o golo**

(Raposo e Miguel, 2013: 717)

Quanto aos sintagmas preposicionais (SP) modificadores do nome, Brito e Raposo (2013) os classificam em dois grupos: os reduzidos e os completos. Os SPs do primeiro grupo são complementados por um SN reduzido (162a), ou seja, sem especificador, ou um verbo no infinitivo (162b). Os SPs do segundo grupo são complementados por um SN completo, ou seja, especificado por um determinante (162c) ou um quantificador (162d)

(Brito e Raposo, 2013: 1067). Os SPs se posicionam sempre à direita do nome com o qual se relacionam, por isso a agramaticalidade das expressões apresentadas a seguir:

(162)
a. livro **de banda desenhada** / ***de banda desenhada** livro
b. máquina **de escrever** / ***de escrever** máquina
c. o homem **da pistola de ouro** / ***da pistola de ouro** o homem
d. o clube **de muitos amigos meus** / * **de muitos amigos meus** o clube

(Brito e Raposo, 2013: 1066-1067, com modificações nossas)

Tendo em vista as características semânticas dos SP classificadores, que influenciam na ordem desses elementos dentro do SN, vale apontar algumas das suas propriedades. Esse tipo de SP têm a função de "classificar o nome modificador relacionando-o com o conjunto de propriedades denotado pelo substantivo ou verbo que é complemento da preposição" (Brito e Raposo, 2013: 1067). Retomando o exemplo (162b), o infinitivo em *máquina de escrever* não "tem qualquer valor referencial; antes representa semanticamente um tipo" (Brito e Raposo, 2013: 1067). O SP em *máquina de escrever* introduz "uma classificação na classe das máquinas, e a expressão denota um subtipo particular destas por oposição a outros, como, p.e., *máquina de filmar*, *máquina de barbear*, *máquina de costura*" (Brito e Raposo, 2013: 1067).

Devido a essa forte relação semântica entre o nome o SP classificador, esse tipo de SP ocorre imediatamente após o nome modificado por ele (Brito e Raposo, 2013: 1107). Caso outro elemento se posicione entre o nome e esse tipo de SP, há um forte estranhamento (163a), principalmente se a combinação "nome + SP" constitui uma **unidade multilexical**, ou seja, "uma expressão em que há uma coesão interna entre os seus elementos constitutivos" (Brito e Raposo, 2013: 1069). A expressão *barco à vela* em (163b) exemplifica esse tipo de unidade multilexical

(163)
a. máquina **de escrever** *antiga* / ?? máquina *antiga* **de escrever**
b. barco **à vela** *ligeiro* / ?? barco *ligeiro* **à vela**

(Brito e Raposo, 2013: 1069)

No caso da ocorrência de um SP com e um SA pós-nominal modificando o mesmo nome, a ordem preferida em português é Nome + SA + SP (Brito e Raposo, 2013: 1107). Muitas vezes, alterando-se essa ordem, o SN pode causar estranhamento (164b).

(164)

a. ... o melhor cantor *romântico* **de sua geração**.
b. ?? o melhor cantor **de sua geração** *romântico.*

(Corpus: CETENFolha, com modificações em (164b))

Brito e Raposo (2013) ressaltam que há casos em que "a ordem nome + adjetivo + sintagma preposicional (classificador) permite evitar possíveis ambiguidades" (Brito e Raposo, 2013: 1069). Nos seguintes excertos, em (165a) o adjetivo *profissional* modifica *campeonato*, já em (165b) se entende que esse adjetivo modifica o nome *basquete* e não *campeonato*, mudando o sentido do SN:

(165)

a. ... o campeonato *profissional* **de basquete** ...
b. ... o campeonato **de basquete** *profissional*...

(Corpus: CETENFolha, com modificações nossas)

Os modificadores oracionais se inserem preferencialmente no final do SN, por isso a agramaticalidade de (166b) e (166c)[23]. Isso ocorre possivelmente devido "a uma tendência para colocar os constituintes mais extensos e complexos na posição final dos sintagmas (ou das frases) em que ocorrem" (Brito e Raposo, 2013: 1108).

(166)

a. ... a estrela *solitária* **que aparece na bandeira do Texas**
b. *... a estrela **que aparece na bandeira do Texas** *solitária*
c. ... a estrela *de cinco pontas* **que aparece na bandeira do Texas**
d. *... a estrela **que aparece na bandeira do Texas** *de cinco pontas*

(Corpus: CETENFolha, com modificações em (166b) e (166d))

Tendo em vista a mobilidade dos SAs dentro do SN, os parágrafos seguintes se dedicam a apresentação das características gerais dos adjetivos em português e a influência da posição pré ou pós-nominal dos SAs modificadores no significado do SN.

Por vezes, a ordem dos SAs altera a interpretação do SN, pois a posição deles pode exercer influência na interpretação [+ ou –] restritiva da expressão nominal. Isto posto, a descrição da ordem dos SAs apresentada a seguir visa examinar essa relação entre a posição e o traço semântico de restrição, com base em Brito e Raposo (2013).

[23] Os exemplos (166b) e (166c) são possíveis em PB, mas são agramaticais na leitura pretendida, ou seja, equivalentes a (166a) e (166c), respectivamente.

Os adjetivos atributivos podem ser divididos em quatro grandes classes com propriedades distintas: os adjetivos qualificativos (167a), os adjetivos relacionais (167b), os adjetivos avaliativos (167c) e os adjetivos intencionais (167d). Os SN apresentados a seguir exemplificam cada um desses adjetivos:

(167)

a. uma mesa **comprida**
b. uma paisagem **campestre**
c. um lugar **agradável**
d. o **falso** culpado

(Brito e Raposo, 2013: 1086-1087)

A classe dos qualificativos corresponde a adjetivos "que introduzem uma propriedade simples nos nomes, no âmbito de um dos parâmetros de natureza física, psicológica ou conceptual constitutivos das entidades que os nomes denotam" (Brito e Raposo, 2013: 1086), como ilustrado nos seguintes SNs em que os parâmetros foram indicados entre parênteses:

(168)

a. uma mesa **comprida** (adjetivo qualificativo de dimensão)
b. uma bandeira **azul** (adjetivo qualificativo de cor)
c. um professor **inteligente** (adjetivo qualificativo de estado psicológico)
d. um aluno **doente** (adjetivo qualificativo de estado fisiológico)

(Brito e Raposo, 2013: 1086)

Em posição pós-nominal, os adjetivos qualificativos possuem interpretação restritiva e são uma peça fundamental na identificação do referente de SN definidos. Por exemplo, em (169a), há a possibilidade de existirem outros rapazes nesse contexto, mas o SA que inclui o adjetivo *doente* permite identificar, de entre esse grupo, qual rapaz é mencionado pela *dona-de-casa*. No caso de SNs indefinidos, como em (169b), os adjetivos qualificativos pós-nominais delimitam o tipo de entidade denotada, mas não são referenciais, pois não auxiliam na identificação do seu referente.

(169)

a. A dona-de-casa disse que o Gapa sabia que a família d**o rapaz doente** é pobre.
b. ...isso torna difícil o socorro a **um índio doente**.

(Corpus: CETENFolha)

É possível que certos adjetivos qualificativos, principalmente os que possuem uma componente avaliativa subjetiva mais forte, ocorram em posição pré-nominal:

(170)

a. ...o segundo governador baiano, em uma década, nascido *na* ***pequena*** *Acajutiba*.
b. *Num* ***raro*** *momento* de descontração entre os jogadores...

(Corpus: CETENFolha)

Nessa posição, esses adjetivos possuem uma interpretação não restritiva, ou seja, não restringem o conjunto denotado pelo nome modificado, comportamento diferente dos adjetivos qualificativos em posição pós-nominal. De acordo com Brito e Raposo (2013),

> "estes adjetivos veiculam uma informação suplementar sobre o referente, realçando ou intensificando uma propriedade deste que não é, no entanto, usada na sua caracterização básica nem serve para a sua identificação pelo ouvinte. Trata-se de uma espécie de comentário do falante que realça e intensifica uma propriedade do referente, esse comentário tem frequentemente uma conotação subjetiva e avaliativa" (Brito e Raposo, 2013: 1090).

De acordo com esses autores, o falante que produz um SN definido não possui a intenção que um adjetivo pré-nominal seja utilizado como informação para identificar o referente. Essa propriedade subjetiva dos adjetivos qualificativos em posição pré-nominal pode explicar porque certos adjetivos qualificativos "que, em posição pós-nominal, denotam propriedades materiais constitutivas das pessoas e têm leitura restritiva, adquiram uma leitura figurada, altamente expressiva, quando ocorrem antes do nome" (Brito e Raposo, 2013: 1091), como nos SN assinalados em (171).

(171)

a. ...tomasse *as* ***velhas*** *personagens* como se fossem fantoches...
b. Mas o presidente é *um* ***grande*** *homem*, porque sua timidez…

(Corpus: CETENFolha)

Fazem parte da classe dos relacionais os adjetivos que surgiram a partir de substantivos através de processos morfológicos ou etimológicos. A função desse tipo de adjetivo é estabelecer uma relação entre dois sentidos: "o sentido do nome de que o adjetivo deriva ou que se liga a ele etimologicamente e o sentido do nome modificado pelo adjetivo" (Brito e Raposo, 2013: 1094). Em (172a) o adjetivo *campestre* deriva do

substantivo "*campo*" e relaciona o sentido denotado por esse substantivo com a palavra "*paisagem*". Em (172b) o adjetivo *fluvial* está ligado com o sentido de "*rio*" e relaciona esse sentido com o nome "*pesca*".

(172)

a. paisagem **campestre**
b. pesca **fluvial**

(Brito e Raposo, 2013: 1086)

Segundo esses autores, "estes adjetivos têm como função classificar o nome modificado em subtipos, de acordo com a relação estabelecida com o sentido do adjetivo" (Brito e Raposo, 2013: 1095). O aspecto restritivo está incluído na própria definição dos adjetivos relacionais, que "não são suscetíveis de ter uma interpretação avaliativa, de natureza subjetiva, o que explica o fato de não poderem ocorrer em posição pré-nominal, com leitura não-restritiva" (Brito e Raposo, 2013: 1095). Consequentemente, não podem ser posicionados antes do nome, como é possível observar nos exemplos (173b) e (174b):

(173)

a. ...podem ser suspensos e *a reserva* ***orçamentária***, cancelada.
b. * ...podem ser suspensos e *a* ***orçamentária*** *reserva*, cancelada.

(174)

a. Alguns parlamentares atribuíram a renúncia a *uma jogada* ***política***...
b. * Alguns parlamentares atribuíram a renúncia a *uma* ***política*** *jogada*...

(Corpus: CETENFolha, com modificações em (173b) e (174b))

A associação "nome + adjetivo relacional" geralmente constitui uma unidade multilexical com alto grau de coesão, por isso outros adjetivos qualificativos ou elementos modificadores restritivos do nome não podem interromper essa adjacência (Brito e Raposo, 2013: 1095), como pode ser verificado nos seguintes exemplos, em que outros adjetivos, (175a) e (175b), e um SP (175a), se inserem entre o nome e o adjetivo relacional:

(175)

a. **ordem com dez membros* ***religiosa***
b. **avião antigo* ***presidencial***
c. **sobremesa deliciosa* ***caseira***

(Brito e Raposo, 2013: 1095)

Quanto aos adjetivos avaliativos, são utilizados pelo falante para introduzir "uma apreciação subjetiva sobre a entidade denotada pelo nome, mas não representam características constitutivas das entidades (...) nem relacionam o nome modificado com o sentido de outro nome" (Brito e Raposo, 2013: 1086), como é possível perceber nos seguintes exemplos (176). Além disso, esses adjetivos "não têm uma interpretação restritiva, quer em posição pré-nominal (..), quer em posição pós-nominal" (Brito e Raposo, 2013: 1092).

(176)

a. um sucesso **fantástico**
b. uma pessoa **agradável**

Para esclarecer o possível estranhamento que adjetivos avaliativos podem causar quando estão na posição pós-nominal em SN definidos, Brito e Raposo (2013: 1092) fazem um paralelo entre esses adjetivos e os qualificativos. De acordo com esses autores, quando o SN é definido, os adjetivos pós-nominais devem auxiliar o ouvinte na identificação do referente. Os adjetivos qualificativos em (177a) e (177b) dão esse auxílio devido a sua propriedade restritiva em posição pós-nominal, mas os adjetivos avaliativos *magnífica* (178a) e *maravilhosa* (178b) podem causar estranhamento nessa posição, pois a sua natureza não restritiva não ajuda na identificação do referente.

(177)

a. Hoje vou pôr *a gravata* ***azul*** (adjetivo qualificativo)
b. Parti *a jarra* ***redonda*** (adjetivo qualificativo)

(178)

a. # Hoje vou pôr *a gravata* ***magnífica*** (adjetivo avaliativo)
b. # Parti *a jarra* ***maravilhosa*** (adjetivo avaliativo)

(Brito e Raposo, 2013: 1093)

Quando esses adjetivos avaliativos são colocados na posição pré-nominal em um SN definido, como exemplifica (179), não há estranhamento, pois o falante não espera uma interpretação restritiva. A sentença (180a) causa estranhamento semântico, pois ambos os adjetivos (avaliativo e qualificativo) ocorrem em posição pós-nominal no SN definido. Contudo, se o adjetivo avaliativo ocorre em posição pré-nominal e o qualificativo ocorre em posição pós-nominal (180b) a frase é aceita pelos leitores. Já no caso de um SN indefinido (180c), em que não se espera que o falante identifique o referente, não há

estranhamento na ocorrência simultânea desses adjetivos em posição pós-nominal, pois ambos podem auxiliar na identificação do referente.

(179)

a. Hoje vou pôr *a* ***magnífica*** *gravata*
b. Parti *a* ***maravilhosa*** *jarra*

(180)

a. # Hoje vou pôr *a gravata azul* ***magnífica***
b. Hoje vou pôr *a* ***magnífica*** *gravata azul*
c. Hoje vou pôr *uma gravata (azul)* ***magnífica***

(Brito e Raposo, 2013: 1093)

Ainda acerca da ordem desses dois tipos de adjetivos no SN, "quando um adjetivo qualificativo ocorre com um adjetivo avaliativo, ambos em posição pós-nominal, o adjetivo avaliativo ocorre preferencialmente depois do adjetivo qualificativo" (Brito e Raposo, 2013: 1107). Isso pode ser observado no exemplo (181), em que o excerto causa estranhamento ao se posicionar o adjetivo avaliativo *deslumbrante* antes do adjetivo qualificativo *verde* (181b)*:*

(181)

a. ...uma ilha *verde* **deslumbrante**, que é São Luís...
b. ?? ...uma ilha **deslumbrante** *verde*, que é São Luís...

(Corpus: CETENFolha, com modificações em (181b))

Para finalizar, os adjetivos intensionais "quando ocorrem em posição pré-nominal, não denotam qualquer propriedade nem restringem a classe denotada pelo nome" (Brito e Raposo, 2013: 1087). Eles também não fazem apresentam uma apreciação subjetiva do falante acerca da entidade expressada pelo nome. Ao invés disso, eles servem para introduzir "uma determinada avaliação (do falante) sobre a aplicabilidade do próprio nome ao referente do sintagma nominal" (Brito e Raposo, 2013: 1087). Fazem parte desse grupo adjetivos como *falso*, *suposto* e *verdadeiro*, em posição pré-nominal, como exemplificados respectivamente:

(182)

a. ...há possibilidade de Aragão ser indiciado por ***falsa*** *notícia* de crime.
b. ...se não temia perder votos por conta de *seu* ***suposto*** *ateísmo*...
c. ...mas é também considerado *uma* ***verdadeira*** *fábrica de dinheiro*.

(Corpus: CETENFolha)

Tomando as ideias de Brito e Raposo (2013), o valor semântico do adjetivo *suposto* em (182b) não é adicionar uma característica ao nome *ateísmo*, nem restringir à classe dos ateístas que são "supostos". Esse adjetivo deixa "em aberto (até ver) a aplicabilidade do sentido do próprio nome (...) ao referente do sintagma nominal" (Brito e Raposo, 2013: 1087). Ao usar o adjetivo *suposto*, o falante está implicitamente informando que a aplicabilidade do nome *ateísmo* à pessoa terá de ser confirmada.

5.2.2.3 Ordem dos complementos

Como já foi assinalado na seção 5.1.1.3, existe uma distinção entre nomes autônomos e nomes dependentes. O que os distingue é, como se viu, o fato de os segundos, mas não os primeiros, selecionarem complementos (ver também Peres, 2013).

Segundo Raposo e Miguel (2013: 715), os argumentos selecionados por nomes dependentes exercem a função gramatical de complemento, ocorrendo na posição à sua direita, e são tipicamente introduzidos por uma preposição, como no excerto assinalado em (183a). Eles também podem ser constituídos por possessivo (183b) ou se manifestarem como argumentos adjetivais (183c).

(183)

a. *A sequência* ***de charges*** nos jornais...
b. ...é *um velho amigo* ***seu***.
c. ...fica uma *invasão* **branca**, inviável...

(Corpus: CETENFolha)

Os autores Brito e Raposo (2013) apontam que "enquanto os argumentos de um verbo são tipicamente obrigatórios (embora nem sempre), os argumentos de um nome, tal como os seus modificadores, são tipicamente opcionais" (Brito e Raposo, 2013: 1061). Essa afirmação pode ser observada nos seguintes exemplos em que os argumentos do nome e os argumentos correspondentes do verbo foram sublinhados, seguindo as ideias de Brito e Raposo (2013: 1048). A frase que corresponde às nominalizações foi representada em (184a). Os argumentos do nome deverbal *chegada* em (184b) correspondem aos argumentos do verbo "*chegar*" (184a). Entretanto, é possível notar que a ausência dos complementos nominais em (184c) não prejudica a gramaticalidade da frase:

(184)

a. soldados do exército **chegaram** aos morros cariocas
b. A **chegada** de soldados do Exército aos morros cariocas tem atraído as adolescentes.
c. A **chegada** tem atraído as adolescentes.

(Corpus: CETENFolha, com grifos e modificações)

Nos SNs que têm como núcleo um nome deverbal (cf. seção 5.1.1.3; Brito e Raposo, 2013: 1049-1050), como nos exemplos (185b-d) o argumento do nome que equivale ao sujeito da frase (*Pedro*) pode ser realizado como pronome possessivo (*seu* em 185b), na forma casual genitiva; ou através de um SP introduzido pela preposição *de* (*do Pedro* em 185c). O argumento de 3ª pessoa também pode ser realizado com pronome na forma casual oblíqua (*ele* em 185d) introduzido pela preposição *de*.

(185)

a. ***Pedro*** *abraçou* a Joaquina.
b. *O* ***seu*** *abraço* à Joaquina.
c. *O abraço* ***do Pedro*** à Joaquina.
d. *O abraço* ***dele*** à Joaquina.

(Brito e Raposo, 2013: 1050, com grifos e modificações)

Em português os nomes não podem reger complementos que são SN, sendo necessária a introdução de uma preposição antes do complemento nominal (Brito e Raposo, 2013: 1051-1052). Por isso, o complemento não preposicionado é agramatical em (186a) e adequado em (186b), após a inserção da preposição *de* entre o núcleo e seu complemento:

(186)

a. *um *desenho* **a cidade**
b. um *desenho* **da cidade**

(Brito e Raposo, 2013: 1051)

Essa inserção obrigatória de preposição no início do complemento nominal também pode ser observada em complementos oracionais infinitivos do nome: "quando a oração é infinitiva, é obrigatoriamente introduzida por uma preposição (...) se a oração não é introduzida por preposição na frase correspondente, a preposição *de* é usada na nominalização" (Brito e Raposo, 2013: 1052). Como é possível observar nos excertos seguintes, ambos os complementos infinitivos foram introduzidos por preposições. O

verbo "*interessar*" requer a preposição *em*, por isso a presença dessa preposição na nominalização desse verbo em (187a); e o verbo "*prometer*" não solicita preposição, por isso a inserção da preposição *de* antes do infinitivo (187b).

(187)

a. ...demonstrado o *interesse* **em *atrair* as organizações não-governamentais**...
b. ...sua *promessa* **de *recuperar* o poder aquisitivo da população.**

(Corpus: CETENFolha)

No caso de o argumento do nome ser uma oração finita, "introduzida pelo complementador *que*, (...) a presença da preposição *de* é bastante frequente, mas não tem o caráter de obrigatoriedade que tem com uma oração infinitiva ou com um sintagma nominal" (Brito e Raposo, 2013: 1052-1053). Isso pode ser observado nos excertos apresentados a seguir em que os nomes *declaração* e *interesse* são complementados por orações finitas, mas em (188a) há a inserção da preposição *de* antes do complementador *que* e em (188b) não há preposição entre o nome e *que*.

(188)

a. A simples *declaração* **de que** haverá um fundo cujos recursos...
b. O governo tem *interesse* **que** os governadores tenham visualização clara destas contas...

(Corpus: CETENFolha)

Acerca da posição dos complementos oracionais, "sempre que a oração completiva ocorre com outros complementos ou modificadores do nome, ocupa, geralmente, a posição final no sintagma nominal" (Barbosa, 2013: 1879). Os SAs e os SPs geralmente precedem esse tipo de oração, como é possível verificar nos exemplos seguintes: o deslocamento do SA (189b) e do SP em (190b) para o final do SN, após os complementos oracionais, causa estranhamento.

(189)

a. uma *explicação* [pouco plausível] de [como o assalto se deu]
b. ?? uma *explicação* de [como o assalto se deu] [pouco plausível]

(190)

a. a *exigência* [dos trabalhadores] de [que os salários sejam aumentados]
b. ?? a *exigência* de [que os salários sejam aumentados] [dos trabalhadores]

(Barbosa, 2013: 1879, com os grifos da autora)

Segundo Barbosa (2013), esse comportamento "deve-se a uma tendência mais geral da língua para colocar os constituintes mais longos e complexos em posição final" (2013: 1879). Apesar disso, essa preferência é neutralizada quando os outros complementos ou modificadores do SN são muito extensos. Tendo em vista a extensão do SA em (191a) e do SP em (191b), a presença desses sintagmas no final do SN não causa estranhamento:

(191)

a. uma explicação de [como o assalto se deu] muitíssimo pouco plausível
b. a exigência de [que os salários sejam aumentados] da parte dos trabalhadores

(Barbosa, 2013: 1879, com os grifos da autora)

6. ANÁLISE DOS DADOS

Este capítulo se dedica à análise dos dados recebidos para a elaboração da presente investigação e será dividido em duas seções. Na primeira (6.1), a partir das *Guidelines* fornecidas pela empresa, será observada nos dados a manifestação de dois tipos de erros relativos ao processo de anotação, nomeadamente os erros de segmentação (6.1.1) e os erros de categorização (6.1.2). Na segunda seção, serão observados com mais atenção os erros de concordância (6.2.1) e ordem de palavras (6.2.2) encontrados nos dados, procurando-se entender a motivação por trás desses erros para auxiliar na construção de sugestões que possam ajudar o processo de anotação feito na empresa.

A anotação é uma etapa de avaliação da qualidade da tradução feita manualmente por anotadores humanos através da categorização dos erros encontrados na tradução final (cf. seção 4.2). No caso da Unbabel, os erros categorizados pelos anotadores passaram por, no mínimo, dois "processos": o sistema de tradução automática e os pós-editores humanos (cf. seção 3.1). Como se verá neste capítulo, alguns erros persistem mesmo após essas duas etapas e os resultados da anotação podem auxiliar a compreender a origem desses erros e a procurar possíveis soluções para os problemas encontrados.

Os dados analisados na presente pesquisa foram anotados sob as etiquetas *Agreement* e *Word Order*[24] pelos anotadores dentro da plataforma da empresa entre os meses de novembro de 2017 e janeiro de 2018. Esses dados possuem o inglês como LP e o PB como LC. Juntamente com os trechos dos textos em que se encontravam os erros anotados, foram também recebidos os segmentos correspondentes ao erro, a severidade apontada pelo anotador e as instruções do cliente para a tradução do trecho. As instruções do cliente auxiliaram na análise, pois forneceram mais informações acerca do contexto discursivo e ajudaram a resolver ambiguidades e dúvidas acerca das entidades mencionadas. Os trechos dos textos em que se encontravam os erros de concordância e de ordem de palavras foram o foco principal da presente análise, pois forneceram informações acerca do contexto semântico e sintático do erro na LC e seu correspondente na LP, possibilitando uma análise mais aprofundada da origem do erro e a proposta de possíveis soluções.

A empresa forneceu 96 dados categorizados sob a etiqueta *Word Order* e 150 dados categorizados sob *Agreement*, totalizando 246 dados. Porém, como se mostrará mais

[24] Neste trabalho, serão mantidos os termos originais em inglês das etiquetas quando elas se referem à tipologia de erros utilizada pela empresa.

adiante, os dados assim etiquetados nem sempre correspondem efetivamente a erros de concordância ou ordem de palavras. Além disso, alguns dos dados não eram válidos para a análise, pois continham informação insuficiente ou foram demasiadamente alterados após a anonimização manual, impossibilitando a sua observação. Por isso, antes da análise propriamente dita, seguiram-se três etapas: anonimização das entidades mencionadas e das informações sensíveis por razões de privacidade; retirada de dados com *informação insuficiente para a análise* (doravante IIA); e retirada de dados repetidos[25].

Após a retirada dos dados IIA e dos dados repetidos, foram considerados como dados válidos para a análise da presente pesquisa 109 dados etiquetados como *Agreement* e 66 etiquetados como *Word Order*, totalizando 175 dados. A tabela 4 resume esses dados:

DADOS VÁLIDOS	**Agreement**	**Word Order**
Dados fornecidos pela Unbabel	150	96
IIA	- 20	- 14
Repetidos retirados das contagens	- 21	- 16
TOTAL	**109**	**66**

Tabela 4 – Dados de *Agreement* e *Word Order* válidos para a análise

6.1 Erros relativos ao processo de anotação

Antes de observar os erros de concordância e de ordem de palavras presentes nas traduções, foi necessário observar os erros de anotação em si mesmos, assinalando-se os casos em que há diferenças na seleção de segmentos e na escolha das tipologias de erros. Através dessa observação, foi possível retirar os dados que não continham erros ou fenômenos de concordância ou de ordem de palavras, permitindo assim a análise somente desses tipos de problemas na seção 6.2. Além disso, foi possível fazer a distinção entre casos em que os anotadores não seguiram as instruções fornecidas pelas *Guidelines* da empresa e casos em que não fica claro qual seria a melhor maneira de anotar, propiciando assim a falta de uniformidade no processo de anotação. A análise dessas ocorrências permitirá a formulação de possíveis soluções que possam auxiliar na uniformização da anotação feita na empresa, melhorando a fiabilidade da ferramenta.

[25] Para mais informações acerca dos processos de anonimização manual e retirada de dados, ver o Anexo 4 da presente pesquisa.

6.1.1 Erros de segmentação

Como já foi apresentado na seção 4.3 e apontado por Nowak e Ruger (2011) e Burchardt e Lommel (2014), obter resultados de anotação uniformes não é tarefa fácil, sendo comum obter uma variação nos resultados de anotação. O aumento da concordância inter-anotadores é um dos principais objetivos das *Annotation Guidelines*, pois essa variação influencia os resultados das anotações. Essas *Guidelines* possuem uma seção inteira dedicada ao processo de segmentação (*unitizing process*) com exemplos de segmentação dos erros mais difíceis de categorizar (cf. seção 4.2.2). Porém, mesmo tendo acesso a essas orientações, os anotadores cometem alguns erros de segmentação que podem influenciar negativamente a uniformidade da anotação e a nota final de qualidade da tradução feita na empresa. Nesta subseção, procura-se observar esses tipos de erros nos dados fornecidos pela empresa, buscando sugestões que possam contribuir para a uniformização da anotação.

A tabela 5 apresentada a seguir demonstra essa variação relativamente aos segmentos inseridos na etiqueta *Agreement* (A28 e A29) e na etiqueta *Word Order* (W54 e W55).

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
A28	*...screenshots of the error message you're seeing.*	*.... imagens de captura de tela [do] mensagem de erro você está vendo.*	do
A29	*Full page screenshot of the error message..*	*Página completa imagem de captura de tela [do] [mensagem] de erro...*	do/mensagem
W54	*It is possible to purchase our DJ products directly...*	*É possível comprar [produtos] do [nosso] [DJ] diretamente...*	produtos/nosso/DJ
W55	*It is possible to purchase our DJ products directly...*	*É possível comprar [produtos] [do] [nosso] [DJ] diretamente...*	produtos/do/nosso/ DJ

Tabela 5 – Variação na segmentação dos dados

Nas primeiras duas linhas, apesar de estar inserido em contextos diferentes, o excerto "*the error message"* na LP foi traduzido incorretamente da mesma maneira em A28 e A29[26]: "*do mensagem de erro*". Nesse caso, há um erro de concordância de gênero entre

[26] Os códigos dos segmentos apresentados na presente seção se referem ao número do dado na listagem original, disponibilizada somente ao júri avaliador: a letra "A" corresponde a exemplos categorizados como

o artigo definido (contraído com a preposição) e o núcleo do SN, o nome *mensagem*. Por um lado, em A28, somente a unidade [do] foi anotada. Por outro lado, em A29, o nome que é núcleo do SN também foi incluído na segmentação, logo as duas unidades [do] e [mensagem] foram selecionadas. Nas duas últimas linhas, os textos na LP e de chegada são idênticos, mas a segmentação foi feita de maneira distinta: três unidades foram selecionadas em W54 e quatro unidades foram selecionadas em W55.

Tendo em vista os dados originalmente fornecidos pela empresa terem sido selecionados sob as etiquetas *Word Order* e *Agreement*, as orientações e exemplos inseridos nesta seção tratam principalmente desses dois tipos de etiquetas. Serão apresentados também alguns exemplos que mostram casos de variação na segmentação dos erros feitos pelos anotadores da Unbabel. Nesta seção, não serão analisados propriamente os erros[27], mas apenas de que modo foi feita a segmentação. Por isso, não será discutida a ocorrência (ou ausência) de erros de concordância ou ordem palavras. Ao invés disso, será dado destaque ao processo de segmentação das unidades, através da indicação das inconsistências encontradas.

6.1.1.1 Segmentação segundo as *Annotation Guidelines*

Para compreender qual dentre as diferentes segmentações apresentadas nos dados fornecidos pela Unbabel é a mais adequada, é necessário observar as orientações das *Annotation Guidelines* acerca da seleção das unidades incorretas. A seleção das unidades com erros envolvendo as categorias de *Agreement* e *Word Order* foram detalhadas com exemplos em duas seções das *Annotation Guidelines*: na seção referente a "*Annotation Rules*" e na seção de "*Tricky Cases*".

Essas *Guidelines* indicam que todas as unidades que incluem o mesmo erro devem ser selecionadas individualmente, depois disso todas essas unidades devem ser categorizadas sob a mesma etiqueta. Ressalta-se ainda que é necessário manter uma marcação mínima (*minimal markup*) e que os erros devem estar contidos no menor intervalo possível (*shortest possible span*). Assim, mesmo se os termos incorretos se encontram separados por outras palavras, somente os constituintes que fazem parte do erro podem ser selecionados. As orientações da empresa resumem esta questão da seguinte maneira:

Agreement pelos anotadores da Unbabel e a letra "W" corresponde a trechos categorizados em *Word Order* pelos mesmos anotadores.

[27] É possível encontrar nesta seção exemplos em que não há efetivamente erro de concordância ou de ordem de palavras, tendo em conta o que foi exposto na seção 5 relativamente à descrição das estruturas a estudar neste relatório. Esses problemas serão discutidos com mais detalhes nas seções 6.2.1 e 6.2.2, dedicadas a essa matéria.

> "If there are two or more units spread apart within the text that form a single error, then you will need to select all the units and choose one type of error. If there are individual errors spread across the text that belong to a same type of error, then you will need to select each error and its type individually" (*Annotation Guidelines*: 11).

No caso da seleção de erros relativos à etiqueta *Agreement*, as *Annotation Guidelines* apontam que esse tipo de erro pode causar dificuldades, pois as partes discordantes podem se encontrar separadas no texto. As orientações da empresa sugerem que seja selecionado "*the minimal spam to fix the issue*", ou seja, o menor intervalo para solucionar o problema. Os seguintes exemplos com erros de concordância verbal em inglês foram apresentados pelas *Guidelines*, nas citadas seções *Annotation Rules* (192) e *Tricky Cases* (193), para auxiliar os anotadores:

(192)

a. She advised that **the amount** charged on your credit card by the hotel **were** a mistake.

(*Annotation Guidelines*: 11, com os grifos originais)

(193)

a. *Translation*: The **man** whom they saw on Friday **were** very big
b. *Correct*: The **man** whom they saw on Friday **was** very big

(*Annotation Guidelines*: 28, com os grifos originais)

Quanto à segmentação, no exemplo (192) foram selecionadas duas unidades: um só bloco contendo o determinante, *the*, e o núcleo, *amount*, do SN sujeito "t*he amount charged on your credit card by the hotel*" e outro bloco com o verbo da oração subordinada [*were*]. No exemplo (193), ambas as orações com a tradução incorreta (193a) e correta (193b) foram apresentadas nas *Guidelines*. Nos dois casos, são selecionados o núcleo, *man*, do SN sujeito "*the man whom they saw on Friday*" e o verbo da oração principal, o que resultou na seleção de [*man*] e [*were*] em (193a); e de [*man*] e [*was*] em (193b).

Como as próprias *Guidelines* salientam, é difícil elaborar uma regra de segmentação e seleção de unidades que seja uniforme e abranja todos os tipos de erros possíveis. Note-se que nem mesmo os exemplos de segmentação de *Agreement* apresentados pelas *Guidelines* são uniformes, uma vez que foram selecionados [Determinante + Núcleo do SN] no exemplo em (192), mas somente o [Núcleo do SN] em (193). Dadas as orientações para selecionar o menor intervalo possível para corrigir o problema e manter uma

marcação mínima, por um lado, a correção da flexão do verbo "*were*" em (193a) resolveria o problema da tradução, não sendo necessário modificar o sujeito da oração; por outro lado somente é possível saber que o verbo está mal flexionado em (193a) ao se observar os traços do núcleo do sujeito "*man*", necessitando assim da seleção desse núcleo.

Quanto aos erros de *Word Order*, as *Annotation Guidelines* apontam que sua segmentação pode ser problemática, pois é necessário identificar as porções de textos que devem ser selecionadas. Essas instruções sugerem que "*you should select the shortest portion of text that could be moved to solve the problem*" (*Annotation Guidelines*: 11). Os erros envolvendo a ordem de palavras são divididos por essas *Guidelines* em dois casos possíveis: as palavras adjacentes (*adjacent words*) e as palavras descontínuas (*discontinuous words*). No primeiro caso, a empresa orienta que ambas as unidades devem ser selecionadas individualmente e depois categorizadas sob a etiqueta *Word Order*, criando assim um único erro em que foi assinalado que ambas as unidades estão na posição incorreta. No caso das palavras descontínuas, as *Guidelines* da empresa afirmam que devem ser selecionados ambos o bloco de palavras e a palavra que precisa ser reordenada.

Para orientar a segmentação dos anotadores quanto aos erros de *Word Order*, as *Guidelines* apresentam exemplos, citados a seguir, nas seções *Annotation Rules* (194) e *Tricky Cases* (195) e (196). Seguindo as orientações da empresa, em (194a), as palavras adjacentes [*lernen*] e [*Deutsch*] devem ser selecionadas individualmente sob a mesma etiqueta *Word Order*; e em (195), as unidades descontínuas [*a small*] e [*only*] devem ser selecionadas. Esse documento também fornece um exemplo de segmentação envolvendo erros de ordem entre palavras classificadas como contíguas (*contiguous words*): em (196), as palavras [*design*] e [*daily*] são selecionadas.

(194) *Adjacent Words*

a. Translation: Du musst schnell **lernen Deutsch**
b. Correct: Du musst schnell **Deutsch lernen**

(*Annotation Guidelines*: 12, com os grifos originais)

(195) *Discontinuous Words*

a. Translation: We use **[a small] [only]** number of the IPs listed on this page
b. Correct: We use **[only] [a small]** number of the IPs listed on this page

(*Annotation Guidelines*: 28, com os grifos originais)

(196) *Contiguous Words*

a. Translation: We also share **design daily** tips on social media
b. Correct: We also share **daily design** tips on social media

(*Annotation Guidelines*: 28, com os grifos originais)

Apesar de apresentar exemplos com esses três tipos de problemas relativos a ordem de palavras, não há uma explicação mais detalhada acerca das diferenças entre eles, das diferentes relações sintáticas manifestadas em cada uma delas e da importância de se conhecer esses conceitos para fazer uma segmentação adequada. Na realidade, não há diferença entre as palavras contíguas e adjacentes, pois ambos conceitos definem palavras que se encontram lado a lado na frase. A distinção interessante para a ordem de palavras que vale destacar nas *Guidelines* seria a diferença entre palavras adjacentes e descontínuas, pois há palavras que são vizinhas (adjacentes), mas pertencem a unidades sintáticas distintas (descontínuas), sendo importante que o editor e o anotador estejam familiarizados com essas noções. Além disso, se consideramos estritamente a sugestão acerca de selecionar somente a menor porção de texto "*that could be moved to solve the problem*", nos três exemplos apresentados somente uma das unidades poderia ter sido selecionada: ao mover [lernen] em (194), [only] em (195) e [design] em (196), cada uma das frases passam a ter a ordem correta.

6.1.1.2 Segmentação encontrada nos dados

Na segmentação dos erros de concordância verbal selecionados pelos anotadores nos dados fornecidos pela empresa, o verbo mal flexionado foi o único elemento selecionado pelo anotador, ao contrário do que é sugerido pelos exemplos (192) e (193) das *Guidelines* apresentados acima. Os excertos inseridos na tabela 6 abaixo exemplificam essa segmentação feita pelos anotadores da empresa: em A94 e em A12, somente os verbos [podem] e [deve], respectivamente, foram selecionados. A estrutura coordenada apresentada em A12 pode levantar dúvidas acerca de qual deveria ser o núcleo nominal a ser selecionado em casos de sujeito nulo, ou seja, como deverá ser feita a anotação do antecedente desse sujeito nulo. Também as estruturas de sujeito composto poderiam ser desafiadoras, pois nesses casos haveria dois núcleos nominais internamente ao sujeito.

Código do segmento	Texto de partida	Texto de chegada	Segmentos selecionados pelos anotadores da Unbabel
A94	*Most visual issues can be resolved..*	*A maioria dos problemas visuais [podem] ser resolvidos...*	podem
A12	*All documents must: • Show the issue date and should not be older than 3 months...*	*Todos os documentos devem: • Mostrar a data de emissão e não [deve] ser superior a 3 meses...*	deve

Tabela 6 – Segmentação de verbos feita pelos anotadores da Unbabel

As *Annotation Guidelines* não apresentam exemplos de concordância nominal, sendo difícil dizer ao certo qual das segmentações em A28 e A29, já apresentadas na tabela 5 e retomadas na tabela 7 a seguir, é a mais adequada. Por um lado, seguindo os exemplos (192) e (193), de concordância verbal, em que são selecionados o verbo com erro de flexão e o núcleo do constituinte que controla a concordância, o exemplo A29 seria o mais adequado pois ambos o determinante e o núcleo do SN, [do] e [mensagem], são selecionados. Por outro lado, seguindo as citadas instruções de manter um "*minimal markup*" e selecionar "*the minimal spam to fix the issue"*, o exemplo A28 seria o mais adequado, pois somente a unidade incorreta, ou seja, o determinante [do], foi anotada.

Código do segmento	Texto de partida	Texto de chegada	Segmentos selecionados pelos anotadores da Unbabel
A28	*...screenshots of the error message you're seeing.*	*.... imagens de captura de tela [do] mensagem de erro você está vendo.*	do
A29	*Full page screenshot of the error message..*	*Página completa imagem de captura de tela [do] [mensagem] de erro...*	do/mensagem

Tabela 7 – Segmentação na concordância nominal

Nos dados de *Word Order* analisados, há casos em que os anotadores seguiram os exemplos das *Annotation Guidelines* quanto à segmentação, como é possível observar na tabela 8: em, W76 as unidades adjacentes [cobrança] e [ingresso] são selecionadas

individualmente; em W89, foram selecionados individualmente, sob a mesma etiqueta *Word Order*, o bloco [possa ajudar] e o clítico [a], adjacentes, mas também descontínuos.

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
W76	*Please send them a billing ticket through URL-0...*	*Por favor enviar eles um [cobrança] [ingresso] através de URL-0...*	cobrança/ingresso
W89	*To help me further assist you, could you kindly provide me...*	*Para que eu [a] [possa ajudar], poderia me fornecer...*	a/possa ajudar

Tabela 8 – Segmentação de palavras adjacentes segundo as orientações da Unbabel

Foram também encontrados casos em que essas orientações não foram observadas. Para exemplificar, em W81, o erro de ordem se encontra nas unidades adjacentes [ícone] e [Configurações] e a troca de lugar entre essas unidades resolveria o problema, mas o elemento [no] também foi selecionado, mesmo não contendo nenhum erro de tradução.

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
W81	*Tap on the Settings icon at the top right corner*	*Toque [no] [Configurações] [ícone] no canto superior direito*	no/ícone/Configurações

Tabela 9 – Segmentação em discordância com os exemplos da Unbabel

É de notar que, em todos os exemplos apresentados pelas *Annotation Guidelines*, as unidades selecionadas se encontram lado a lado na frase e uma simples troca de lugar entre elas resolveria o problema de ordem de palavras. Nesse tipo de contexto, também nos dados analisados foram selecionadas sempre mais de uma unidade, como exemplificam W76 e W89 na tabela 9, apresentada anteriormente. A situação fica ainda mais complexa quando os termos selecionados não são adjacentes: as *Guidelines* não apresentam nenhum exemplo de *Word Order* com casos em que a correção da frase não pode ser feita através de uma troca de lugar entre termos vizinhos. Esse último tipo de

problema na ordem de palavras foi encontrado em alguns dados, que foram segmentados de diversas formas e que serão apresentadas sucintamente nos parágrafos seguintes.

Na tabela 10, em W38, o nome "*fornecedor*" é complementado por "*passagem aérea*", logo esse núcleo nominal deveria estar posicionado antes desse complemento e a preposição "de" deveria ter sido inserida entre esses elementos, mas nesse caso foram selecionadas pelos anotadores da Unbabel as unidades [com] e [fornecedor], mesmo não existindo erro de ordem relativamente à posição da preposição "*com*".

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
W38	*...you have made a booking with a flight provider...*	*...uma reserva [com] passagem aérea [fornecedor]...*	com/fornecedor

Tabela 10 – Exemplo de segmentação (*Word Order*)

A seleção de termos que se encontram bem posicionados na frase parece indicar uma tendência nos anotadores para selecionar a unidade na posição incorreta ou não-natural e o termo próximo do local em que essa unidade deveria estar. Nos exemplos W30 e W70, na tabela 11, as frases soariam mais naturais em PB se as unidades [às vezes] e [legalmente] estivessem à direita dos termos [pois] e [tem], respectivamente, apesar de a ordem no texto de chegada não resultar em sequências agramaticais.

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
W30	*Clear your browser history as accumulated cookies can cause trouble sometimes.*	*Limpe o histórico do seu navegador, [pois] cookies acumulados podem causar problemas [às vezes].*	pois /às vezes
W70	*...Chargebacks can take up to 45 days to get resolved once filed as the other side has 40 days legally to respond to bank inquiries.*	*...as cobranças podem demorar até 45 dias para serem resolvidas uma vez arquivado, pois o outro lado [tem] 40 dias [legalmente] para responder ao banco inquéritos.*	legalmente/tem

Tabela 11 – Exemplos de tendências na segmentação (*Word Order*)

Ao observar os elementos anotados, é possível distinguir outra tendência na segmentação de *Word Order*: há casos em que a ordem da anotação das unidades parece estar ligada à ordem desejada dos elementos na LC. Para exemplificar, em W99, na tabela 12, a ordem correta na LC parece ter sido representada através da ordem de anotação dos segmentos, pois a ordem das unidades anotadas *diretamente*/*com*/ *fornecedor*/*passagem* corresponde à ordem esperada desses elementos em PB: "diretamente com o fornecedor de passagem aérea", tendo o anotador omitido a palavra "*aérea*". O mesmo ocorre na segmentação de W77: a ordem de anotação das unidades *mais*/*uma*/*vez* é a ordem considerada mais natural na LC. Também em W93 a ordem de anotação das unidades corresponde à ordem mais natural dos elementos em PB ("começar do início o processo"), mas tal como em W99, alguns termos foram omitidos pelo anotador, nomeadamente os elementos "*o*", "*de*", "*se*" e "*inscrever*".

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
W99	*We would recommend contacting the flight provider directly...*	Recomendamos entrar em contato [com] [passagem] aérea [fornecedor] [diretamente]...	diretamente /com/fornecedor/ passagem
W77	*... and send verification request once again."*	*... envie um pedido de confirmação [uma] [vez] [mais]."*	mais/uma/vez
W93	*... so you should be able to start with the sign up process from the scrach.*	*... então você deve [começar] o [processo] de se inscrever [do] [início].*	começar/do/início/ /processo

Tabela 12 – Exemplos de tendências na segmentação (*Word Order*)

Essa segmentação é muito problemática, pois insere elementos de constituintes sintáticos distintos no mesmo erro, indo contra aquilo que foi prescrito pela empresa. No exemplo W99 da tabela 12, apesar de [diretamente] e [fornecedor] manifestarem erro de ordem, o termo "fornecedor" faz parte do SN complemento da preposição *com*, e o termo "diretamente" é núcleo de um sintagma adverbial que modifica o SV "entrar em contato com o fornecedor de passagem aérea"; logo, as duas palavras anotadas fazem parte de constituintes sintáticos distintos e deveriam ter sido anotados individualmente sob duas etiquetas de *Word Order* também distintas. O mesmo ocorre em W93, que junta na

mesma etiqueta de *Word Order* elementos que fazem parte do SN, "processo", e do SV, "começar" e "do início", ou seja, de unidades sintáticas distintas.

A segmentação dos clíticos é o caso que mais revela a variação e o incumprimento, por parte dos anotadores, das regras prescritas pelas *Annotation Guidelines*. Essas orientações estabelecem claramente que, se o clítico apresentar erro nos casos de estruturas que envolvem verbos e clíticos separados por hífen, somente o clítico deve ser selecionado (*Annotation Guidelines*: 27), como exemplifica a unidade [lo] anotada em A61, na tabela 13, que foi corretamente anotada segundo as orientações da empresa. Contudo, nos dados analisados, os clíticos foram anotados de maneiras variadas, algumas vezes indo contra o que foi prescrito: em W59, o clítico [lhe] foi selecionado juntamente com o verbo [dar] e, em A71, o clítico [lo] com erro de concordância foi anotado juntamente com o nome que é retomado por esse clítico [PERSON'S NAME F].

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
A61	*...you may also use it to purchase the element.*	*...você também pode usá-[lo] para comprar o elemento.*	lo
W59	*...so we can give you a temporary passcode for the app.*	*...para que possamos [dar]-[lhe] uma senha temporária para o aplicativo.*	lhe /dar
A71	*... so we can help you over the phone.*	*... para que possamos ajudá-[lo] por telefone.*	lo/(PERSON'S NAME F)

Tabela 13 – Exemplos de segmentação de clíticos (*Agreement* e *Word Order*).

Esses exemplos demonstram a irregularidade na segmentação dos dados analisados, em alguns casos não respeitando as orientações prescritas pela empresa, apesar de as *Guidelines* incluírem exemplos e explicações detalhadas. A empresa também solicita explicitamente a atenção dos anotadores através de frases como "please keep this in mind when annotating, as we have been noticing that Annotators are not always complying with this rule when required" (*Annotation Guidelines*: 10), mas ainda é possível encontrar certos erros de segmentação nos dados fornecidos.

6.1.2 Erros de categorização

A presente subseção visa retirar os dados que não apresentam nenhum erro ou que apresentam outro tipo de erro que não se inclui nas categorias de concordância ou de ordem de palavras. Durante esse processo de filtragem, salientam-se as confusões entre etiquetas diferentes e os erros mais comuns feitos pelos anotadores relativamente à categorização. Incluem-se também nos objetivos da presente subseção a observação dos principais erros de categorização feitos pelos anotadores e a listagem das categorias de erros mais confundidas com *Word Order* e *Agreement*. Apesar dos problemas encontrados na segmentação feita pelos anotadores, já apresentados na subseção anterior (6.1.1), optou-se por fazer a análise e a anotação estritamente[28] das unidades selecionadas originalmente pelos anotadores da Unbabel.

A anotação realizada no âmbito da presente pesquisa foi feita manualmente, utilizando-se as *Annotation Guidelines* fornecidas pela empresa, bem como as ideias apresentadas em trabalhos anteriores que também analisaram dados de anotação da empresa Unbabel, nomeadamente as pesquisas de Figueira (2018), Testa (2018) e Comparin (2016). Procurou-se fazer uma anotação o mais objetiva possível, apresentando-se nesta subseção as dificuldades e imprecisões encontradas durante o processo de anotação dos dados.

Alguns trabalhos decorrentes de outras investigações feitas na empresa já apontaram a possibilidade de confusões por parte dos anotadores: Testa (2018: 40) discute a questão da *concordância inter-anotadores* e mostra que há opiniões diferentes entre anotadores (2018: 40). Essa autora estabeleceu critérios de anotação para aumentar a concordância entre anotadores e resolver certas dúvidas, como, por exemplo, a diferenciação entre erros de *Tense/Mood/Aspect* e *Agreement* (Testa, 2018: 47-48). Já Figueira (2018: 13-14) separa as categorias de erros da tipologia utilizada pela Unbabel em três tipos: categorias não ambíguas, categorias ambíguas e categorias aparentemente ambíguas[29]. Segundo o autor, a maior dificuldade do processo de anotação reside nessas duas últimas categorias.

[28] Nos dados apresentados nesta subseção, podem ocorrer inúmeros erros em outros termos da mesma frase em que se encontra o erro de *Agreement* ou *Word Order*. Tendo em vista os objetivos da análise levada a cabo, esses erros não serão contabilizados na presente subseção e serão citados somente quando relevantes para explicitar a anotação das unidades originalmente selecionadas. Caso os erros encontrados nas outras palavras não segmentadas originalmente envolvam fenômenos interessantes para a discussão de ordem de palavras e concordância, a contabilização e análise dessas palavras serão feitas na seção 6.2.

[29] Segundo Figueira (2018), a distinção entre esses grupos pode variar segundo o erro e as línguas envolvidas, mas, para exemplificar, de maneira geral os erros de *Capitalization* estão inseridos no grupo de categorias não ambíguas, os erros de *Lexical Selection* estão no grupo de categorias ambíguas e os erros de *Spelling* são aparentemente ambíguos, pois há delimitações nas *Guideline*s quanto ao seu uso. Para mais informações quanto á definição dessas categorias, cf. Anexo 3.

No caso de erros caracterizados por categorias ambíguas, de fato há a possibilidade de se selecionar duas ou mais categorias para o mesmo erro, pois ambas podem caracterizá-lo. No caso de categorias aparentemente ambíguas, apesar de inicialmente parecer que duas ou mais categorias podem ser selecionadas, as *Annotation Guidelines* da empresa restringem o uso desse tipo de categoria, através de suas instruções e exemplos[30].

Apesar de não serem citadas no decorrer da presente seção, devido à dimensão e aos objetivos da mesma, a divisão das categorias feita por Figueira (2018) foi considerada durante o processo de anotação e análise dos erros. A partir da proposta do autor, foi possível perceber a dificuldade em escolher uma só categoria para a anotação do erro, devido à ambiguidade de certas etiquetas. As dúvidas quanto à categorização no caso de categorias ambíguas ou de categorias aparentemente ambíguas serão problematizadas quando consideradas interessante para os objetivos da presente seção. Quando necessário, também serão apresentadas as definições e as instruções acerca da tipologia de erros apresentada nas *Annotation Guidelines*.

A versão[31] das *Annotation Guidelines* utilizada pelos anotadores cujos dados são analisados na presente pesquisa não permite a seleção de duas categorias para o mesmo termo. Nesse caso, a empresa estabelece que seja selecionado somente o erro mais severo e específico, ou seja, "the one that most affect the quality of the translation)" (*Annotation Guidelines*: 26). Entretanto, na anotação feita manualmente durante a presente pesquisa, foram encontrados os casos em que mais de um erro ocorre no mesmo segmento, sendo então preferível a seleção de várias etiquetas para que o erro seja adequadamente definido e categorizado. Tendo em vista os objetivos da presente seção, durante a análise de categorização, para cada segmento foi feita a enumeração das etiquetas que poderiam delimitar os erros apresentados, não tendo sido feita a escolha do erro mais severo.

A Unbabel demonstra uma preocupação com a naturalidade do texto de chegada, pois ressalta que devem ser evitadas as estruturas que não soam naturais ou que estão demasiadamente próximas do texto original (cf. *Language Guidelines*) e a fluência geral os textos anotados na ferramenta devem ser classificados numa escala de 1 a 5 (cf. *Annotation Guidelines*: 9). Porém, em certos casos que envolvem ordem de palavras ou

[30] Para exemplificar, em *Spelling*, apesar da sua definição citar a discordância de traços entre plural e singular, as *Guidelines* fazem uma distinção entre os erros relacionados com *Spelling* e *Agreement* (Para a definição dessas etiquetas, cf. Tabelas 3.1 e 3.2 no Anexo 3; também cf. a *decision tree* proposta em Figueira, 2018).

[31] É de notar que a versão das *Annotation Guidelines* citada na presente pesquisa e utilizada pelos anotadores da Unbabel já foi atualizada pela empresa antes da publicação da presente pesquisa. Na nova versão, já são aceitas as anotações de até duas etiquetas por palavra, como apontado no trabalho de Figueira (2018: 11).

concordância, é difícil estabelecer se uma estrutura está ou não errada: algumas palavras possuem diversas possibilidades de posicionamento na frase (seção 5.2.1) e há fenômenos de variação no uso da concordância em certas estruturas do português, como nos SV com infinitivo flexionado e nos casos de concordância semântica (seção 5.1.2). Além disso, o PB estar passando por diversos processos de variação, ocorrendo muitas vezes conflitos entre aquilo que é prescrito pela variedade de prestígio da língua e o uso efetivo na língua escrita, principalmente em casos como a posição dos clíticos (cf. seção 5.2.1).

A categorização dos dados que estão gramaticalmente corretos mas que ainda assim suscitam dúvidas devido às especificidades da concordância e da ordem de palavras em PB será discutida na seção 6.2, em que será dada atenção especial às relações sintáticas e semânticas envolvidas nesses casos. Na presente subseção esses tipos de casos foram assinalados com as expressões "Envolvendo concordância semântica", "Variação em PB" e "Possivelmente não natural". Também foi adicionada a expressão "Palavra Estrangeira" para os casos que envolvem palavras que, por terem sido mantidas em inglês no texto de chegada, causam certa dificuldade de tradução.

Finalmente, esta subseção foi subdividida em duas grandes partes: na primeira (6.1.2.1) serão observados os dados válidos com segmentos originalmente anotados em *Agreement*; na segunda (6.1.2.2) serão observados os dados válidos com segmentos originalmente anotados em *Word Order*; e na terceira (6.1.2.3) será feita uma síntese dessas análises.

6.1.2.1 Erros de categorização em *Agreement*

Segundo as *Annotation Guidelines* da empresa, a etiqueta *Agreement* deve ser selecionada quando uma ou mais palavras possui(em) erro de concordância quanto aos traços de número, pessoa, gênero e caso. Na ferramenta de anotação da Unbabel, essa categoria se encontra dentro de uma tipologia de erros com vários níveis hierárquicos (cf. Anexo 3), como é possível observar na figura 11:

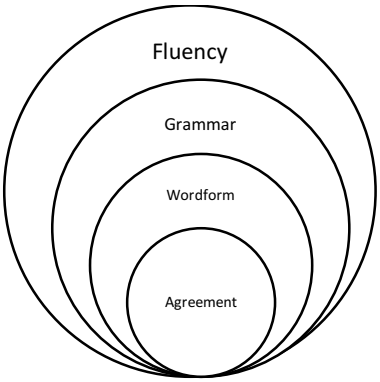

Figura 11 – Níveis hierárquicos envolvendo a etiqueta de *Agreement*

Os próximos parágrafos se dedicam à observação da categorização dos 109 dados válidos originalmente anotados sob a etiqueta de *Agreement*. A partir da observação e (re)anotação das unidades segmentadas, os dados foram separados em três partes distintas: os dados em que há erro de *Agreement* (6.1.2.1.1); os dados e que não há erro (6.1.2.1.2); os dados em que há outro tipo de erro (6.1.2.1.3).

6.1.2.1.1 Os dados em que há erro de *Agreement*

Foram encontrados 79 dados nos quais as unidades selecionadas pelos anotadores efetivamente apresentavam somente erro de concordância, segundo a definição dada pela Unbabel em suas *Annotation Guidelines*. O exemplo apresentado na tabela a seguir mostra esse tipo de dado: em A5 o quantificador [todos] foi selecionado por não concordar em número com o núcleo do SN.

Categorias sugeridas na presente análise	**Nº de dados inseridos nesta categoria**	**Exemplo**			
		Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmento selecionado pelos anotadores da Unbabel**
Agreement	79	A4	*...to all the great content...*	*...a [todos] o grande conteúdo...*	todos

Tabela 14 – Exemplo de categorização: *Agreement*

Também foram encontrados dados em que o erro de *Agreement* ocorreu juntamente com outro erro associado a outras categorias: *Agreement*/*Word Order;*

Agreement/*Diacritics*; *Agreement* /*Wrong Determiner*; e *Agreement*/*Over-Translation* (cf. Tabela 5.1 no Anexo 5). O excerto inserido na tabela 15 abaixo exemplifica o primeiro caso: em A141, o adjetivo [secreto] não concorda em número com o núcleo do SN "*painéis"* e deveria estar após esse último, por isso há simultaneamente erro de ordem e concordância.

Categorias sugeridas na presente análise	**Exemplo**			
	Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
Agreement / Word Order	A141	*...including secret boards because...*	*... incluindo [secreto] painéis...*	secreto

Tabela 15 – Exemplo de categorização: *Agreement* / *Word Order*

O dado apresentado na tabela 16 a seguir exemplifica um erro envolvendo a etiqueta *Agreement* juntamente com mais de uma etiqueta (para mais exemplos, cf. Tabela 5.2, Anexo 5): *Othography* e *Punctuation*[32]. Nesse caso, a correção das unidades anotadas não envolve a mudança dos traços morfossintáticos pré-existentes, mas sim a inserção de parênteses, que modificariam drasticamente a grafia da palavra: devido às características do português, as possíveis variantes de palavras quanto aos traços de número (singular/plural), gênero (feminino/masculino) e até pessoa no caso dos verbos devem ser repetidas entre parênteses "( )" em todos os elementos que estabelecem relações de concordância. Para apresentarem a devida alternância entre a forma singular e plural, a forma "*(ns)*" deveria ter sido inserida no final da palavra "*Imagem*" em A3, a forma "(s)" deveria ter sido inserida após o determinante "a" em A101 e a forma "*(ão)*" deveria ter sido inserida após o verbo "*será*" em A103.

[32] Considerando-se que as *Annotation Guidelines* inserem os parênteses na lista de erros de *Punctuation* (cf. Anexo 3).

Categorias	Nº de dados inseridos nestas categorias	Exemplos				
		Código do segmento	Texto de partida	Texto de chegada	Segmentos selecionados pelos anotadores da Unbabel	Propostas de tradução
Agreement / Orthography / Punctuation	1	A3	*(5) Screenshot(s) of the issue are very helpful!*	*(5) [Imagem] De Captura De Tela(s) da questão são muito úteis!*	Imagem	(5) Imagem(ns) de captura de tela da questão...

Tabela 16 – Exemplo de categorização: *Agreement /Orthography/Word Order*

Foram encontrados 4 dados que envolviam o fenômeno de concordância semântica em português, assinalados com a expressão "*Envolvendo concordância semântica*" na presente pesquisa. Para exemplificar, no dado A94 existem duas possibilidades de flexão do verbo "podem" em PB, apresentados na última coluna: na perspectiva mais convencional, o verbo deve concordar em pessoa e número com o núcleo do sujeito singular "maioria"; contudo, o verbo também poderia se flexionar segundo sentido plural expresso pelo sujeito. Devido à complexidade desses casos, sobre os quais não há uma solução incontestável para a sua classificação (ou não) como erro de concordância, serão dados detalhes na seção 6.2.1.

Fenômeno	Nº de dados inseridos nesta categoria	Exemplos				
		Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Propostas de tradução**
Envolvendo concordância semântica	4	A94	*Most visual issues can be resolved…*	*A maioria dos problemas visuais [podem] ser resolvidos...*	podem	A maioria dos problemas visuais podem ser resolvidos... OU A maioria dos problemas visuais pode ser resolvida...

Tabela 17 – Exemplo de fenômeno problemático: *Envolvendo concordância semântica*

Também foram separados dos outros dados e assinalados através da expressão "Palavra estrangeira" os casos em que havia erro de concordância, mas a palavra foi mantida em inglês no texto de chegada, dificultando assim a definição dos traços de gênero da palavra. O dado A1 apresentado na tabela 18 exemplifica esse tipo de caso: a expressão "*Venture Builder*" é feminina em PB, logo o determinante [um] deveria estar no feminino.

Fenômeno	Exemplos				
	Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Propostas de tradução**
Palavra Estrangeira	A1	*…your company is not a Venture Builder…*	*...sua empresa não é [um] Venture Builder...*	um	... sua empresa não é uma *Venture Builder*...

Tabela 18 – Exemplo de fenômeno problemático: *Palavra Estrangeira*

6.1.2.1.2 Os dados em que "Não há erro"

Considerando somente as unidades selecionadas pelos anotadores, ocorreram 11 casos, originalmente assinalados como erros de *Agreement* pelos anotadores da Unbabel, mas em que os segmentos anotados não continham nenhum tipo de erro. Nesse caso, os dados foram assinalados com a expressão "Não há erro" para efeitos da presente pesquisa. A expressão "Cumprimentos" em A26, apresentada na tabela 19, exemplifica esse tipo de caso. O dado A5, apresentado na mesma tabela, é um caso interessante, pois o segmento [um novo] concorda corretamente com os traços de gênero e número do núcleo do SN "plano", que está na posição incorreta no texto de chegada. Logo, também não há erro nas unidades anotadas, mas sim no termo "plano".

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
A26	*Cheers,*	*[Cumprimentos],*	Cumprimentos
A5	*...a new subscription plan...*	*... [um novo] assinatura plano...*	um novo

Tabela 19 – Exemplos em que "Não há erro" (*Agreement*)

6.1.2.1.3 Os dados em que há outro tipo de erro

No total, foram encontrados 5 erros de categorização envolvendo outras categorias inadequadamente anotadas como *Agreement*, como exemplificado no dado na tabela 20

a seguir: em A64, houve a adição desnecessária do determinante "a" na tradução, ocorrendo *Addition* (para mais exemplos, cf. Tabela 5.3 no Anexo 5).

Categoria	**Exemplo**			
	Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
Addition	A64	*...take a look at our (PRODUCT M) help center article...*	*... olhe a nossa Artigo [da] ajuda da engrenagem de (PRODUCT M)...*	da

Tabela 20 – Exemplo de categorização: *Addition*

Para finalizar, tendo em conta a análise dos 109 dados válidos anotados originalmente com a etiqueta *Agreement*, é possível afirmar que 93 dos dados envolviam efetivamente concordância: 79 deles envolviam somente a etiqueta *Agreement*; 8 envolviam *Agreement* juntamente com outras etiquetas; 4 dados envolvem o fenômeno de concordância semântica; e 2 envolvem uma palavra estrangeira. Os fenômenos de concordância envolvidas nesses erros serão observadas com mais detalhes na seção 6.2.

DADOS ENVOLVENDO FENÔMENOS DE CONCORDÂNCIA	
Erros somente de *Agreement*	79
Erros de *Agreement* com outra(s) etiqueta(s)	8
Dados "Envolvendo concordância semântica"	4
Dados envolvendo "Palavra Estrangeira"	2
TOTAL	**93**

Tabela 21 – Dados válidos envolvendo fenômenos de concordância

6.1.2.2 Erros de categorização em *Word Order*

Segundo a definição das *Annotation Guidelines*, devem ser anotados com a etiqueta *Word Order* os erros que envolvem problemas de ordenação linear no texto de chegada. Os níveis hierárquicos relacionados com essa etiqueta foram representados na figura 12 (cf. também Anexo 3):

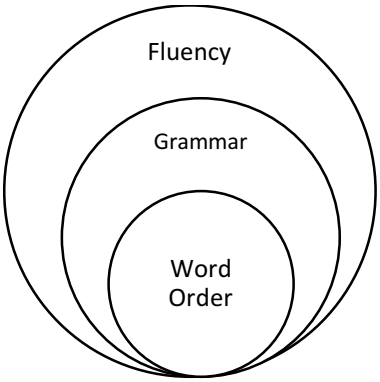

Figura 12 – Níveis hierárquicos envolvendo a etiqueta de *Word Order*

As subseções que se seguem procuram observar a categorização dos 66 dados válidos anotados em *Word Order* pelos anotadores da Unbabel, o que permite verificar os problemas introduzidos pelos anotadores e confirmar algumas dificuldades de categorização quando se usa a tipologia de erros da empresa. Como na subseção anterior, após a (re)anotação manual feita durante esta pesquisa, os dados foram divididos em três grupos: dados em que há erro ou casos problemáticos envolvendo *Word Order* (6.1.2.2.1); dados em que não há erro (6.1.2.2.2); dados em que há outros tipos de erro (6.1.2.2.3).

6.1.2.2.1 Os dados em que há erro de *Word Order*

Dentre os 66 dados válidos originalmente anotados sob a etiqueta *Word Order*, 8 contêm somente erro de ordem de palavras, correspondendo efetivamente à etiqueta atribuída. O dado W100, apresentado na tabela 22 a seguir, exemplifica um caso desse tipo: "*Remove watermark*" é na verdade um nome próprio, designando a ferramenta citada, logo a ordem em PB deve ser "*Remover marca d'água*".

Categoria	**Nº de dados inseridos nesta categoria**	**Exemplo**			
		Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
Word Order	8	W100	*you use the Remove watermark feature*	*...você usa o recurso [marca] [d'água] [Remover]...*	Remover/marca/d'água

Tabela 22 – Exemplo de categorização: *Word Order*

Também foram encontrados dados que envolviam *Word Order* e outras etiquetas. Para exemplificar, no dado W76 da tabela 23, além da ordem incorreta das palavras [configurações] e [cobrança], verifica-se também a falta de preposição precedendo o modificador nominal "ingresso". Apesar de essa omissão da preposição se relacionar mais às características particulares do PB, pois não há preposição no texto original em inglês, as *Annotation Guidelines* destacam que as categorias inseridas na categoria-filha *Omission* também devem ser selecionadas caso o conteúdo esteja faltando somente no texto de chegada.

Categoria	**Exemplo**				
	Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Proposta de tradução**
Word Order / Omitted Preposition	W76	*Please send them a billing ticket...*	*Por favor enviar eles um [cobrança] [ingresso]...*	cobrança/ ingresso	...um ingresso de cobrança...

Tabela 23 – Exemplo de categorização: *Word Order* / *Omitted Preposition*

Devido à segmentação confusa feita pelos anotadores da Unbabel, foram encontrados casos em que unidades com erros distintos foram inseridas sob a mesma etiqueta. Durante a anotação feita para a presente pesquisa, esse fato levou à inserção de diversas etiquetas no mesmo dado, mesmo quando somente uma das palavras selecionadas apresenta esse tipo de erro. Para exemplificar, no dado W25 da tabela 24 apresentada a seguir, foram selecionadas quatro categorias de erros: *Word Order*/*Agreement*/*Omitted Preposition*/*Spelling*. Porém, o erro de *Spelling* se aplica somente à unidade [duração] que estava no plural na LP e deveria também estar no plural na LC.

Categorias	Exemplo				
	Código do segmen-to	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Propostas de tradução**
Word Order / Agreement / Omitted Preposition / Spelling	W25	*... workout minimum/ maximum durations...*	*...[treino] [mínimo] / [duração] [máxima]...*	duração /máxima/ mínimo/ treino	...durações mínimas e máximas de treino... OU ...durações mínimas/ máximas de treino...

Tabela 24 – Exemplo de categorização: *Word Order* / *Agreement* / *Omitted Preposition* / *Spelling*

Vale notar que as categorias que mais coocorrem com a etiqueta *Word Order* são *Omitted Preposition*, *Overly Literal* e *Agreement* (cf. nos exemplos das Tabelas 5.4 e 5.5 do Anexo 5). Estas ocorrências de erros estão relacionadas com as diferentes características do PB e do inglês no que diz respeito a ordem de palavras, as quais serão analisadas na seção 6.2.

O exemplo apresentado na tabela 25 a seguir corrobora a subjetividade no processo de anotação, o que dificultando a uniformidade dos resultados. Em W71, a escolha das categorias depende do olhar do anotador, pois há duas possibilidades diferentes de traduzir o modificador "*bank*": inserido num SN dentro de um SP, como em "aos inquéritos **do banco**", ou inserido num SA dentro de um SN, como em "aos inquéritos **bancários**". No caso da anotação desse erro, é possível selecionar *Word Order* / *Omitted Preposition* / *Omitted Determiner*, caso o anotador considere a estrutura com o SP, ou *Word Order* / *POS* / *Agreement*, caso ele considere a estrutura com SA.

Categorias	Exemplo				
	Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Propostas de tradução**
Word Order / Omitted Preposition / Omitted Determiner **OU** Word Order / POS / Agreement	W71	*...to respond to bank inquiries.*	*... responder ao [banco] [inquéritos].*	inquéritos/ banco	...responder aos inquéritos do banco... **OU** ...responder aos inquéritos bancários...

Tabela 25 – Exemplo de erro com mais de uma possibilidade de tradução (*Word Order*)

Também foram encontrados 26 dados em que o processo de anotação se revela desafiador, pois alguns deles não são erros gramaticais, mas podem ser considerados como problemas de estilo ou traduções demasiadamente literais. Esses casos envolvem geralmente a ordem de estruturas com advérbios ou clíticos em PB, assinalados respectivamente com as expressões “Possivelmente não natural” e “Variação em PB”.

Categorias	Nº de dados inseridos nestas categorias	Exemplo				
		Código do segmento	Texto de partida	Texto de chegada	Segmentos selecionados pelos anotadores da Unbabel	Propostas de tradução
Variação em PB	7	W74	*Please let me know if I can...*	*Por favor, [informe]-[me] se posso...*	me/ informe	...informe-me.../...me informe...
Possivel-mente não natural	19	W75	*...you won't see any data if no one's saved anything from your website yet.*	*...não irá ver nenhum dado se [não] [tiver] guardado qualquer coisa do seu site [ainda].*	ainda /não/ tiver	...se ninguém **ainda** não tiver guardado qualquer coisa do seu site. /...se ninguém não tiver guardado qualquer coisa do seu site **ainda**. /...se ninguém não tiver guardado **ainda** qualquer coisa do seu site.

Tabela 26 – Exemplos de fenômenos problemáticos: *Variação em PB* e *Possivelmente não natural*

É de notar que há também 4 dados que envolvem o fenômeno *Possivelmente não natural* e outra(s) etiqueta(s) da tipologia da empresa (para mais exemplos, cf. Tabela 5.6 do Anexo 5). Esse tipo de dado foi ilustrado na tabela 27: em W63, além da ordem possivelmente não natural de [Diretamente], há o uso inadequado de pontuação e inserção de letra maiúscula.

Categorias	Exemplo				
	Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Propostas de tradução**
Possivel-mente não natural / Punctuation / Diacritics	W63	*...to pay for the order by a (CARD BRAND)/(CARD BRAND) bank card directly.*	*... pagar a encomenda [através] de um cartão bancário (CARD BRAND) / (CARD BRAND). [Diretamente].*	Diretamente/ através	...pagar pela encomenda diretamente através de um cartão bancário... / ...pagar diretamente pela encomenda através de um cartão bancário... / ...pagar pela encomenda através de um cartão bancário diretamente...

Tabela 27 – Exemplo de categorização: *Possivelmente não natural/ Punctuation/ Diacritics*

6.1.2.2.2 Os dados em que "Não há erro"

Ao se considerar estritamente as unidades selecionadas pelos anotadores da Unbabel, foram encontrados 12 dados em que não há nenhum erro. Para exemplificar, no dado W72 apresentado na tabela 28 a seguir os segmentos [link] [de] e [verificação] não foram mal traduzidos na LC.

Código do segmento	Texto de partida	Texto de chegada	Segmentos selecionados pelos anotadores da Unbabel
W72	*Please click the link in your verification email to finish validating your account.*	*Clique no [link] no seu e-mail [de] [verificação] para terminar a validação da sua conta.*	link/de/verificação

Tabela 28 – Exemplo de dado em que não há erro (*Word Order*)

6.1.2.2.3 Os dados em que há outros tipos de erro

Entre os segmentos originalmente anotados, certos dados contêm um só erro que não se enquadra na definição da etiqueta *Word Order*. O exemplo apresentado a seguir ilustra esse tipo de dado: em W23, há a duplicação da palavra "apenas" em dois lugares distintos.

Categoria	Exemplo				
	Código do segmen-to	Texto de partida	Texto de chegada	Segmen-tos seleciona-dos pelos anotado-res da Unbabel	Propostas de tradução
Duplication	W23	*...sensitive information we can provide this only over the phone.*	*...informação sensível, [apenas] [podemos] [fornecer] [isso] apenas telefone.*	podemos /fornecer/ isso/apen as	...apenas podemos fornecer isso por telefone.

Tabela 29 – Exemplos de categorização: *Duplication*

Também foram encontrados dados em que os segmentos apresentavam dois ou mais erros que não se encaixavam como erro de ordem (para mais exempos, cf. Tabela 5.7 no Anexo 5). Para exemplificar, no dado W22 da tabela 30 a seguir, o termo "levantar" foi traduzido de maneira próxima do original, não sendo muito natural o uso desse verbo nesse contexto em PB. Além disso, nesse mesmo dado, a preposição mais adequada para

esse contexto seria [da], pois ela resolveria o problema de ambiguidade presente no texto de chegada: "*um pedido de pagamento da nossa equipe de finanças para você*".

Categoria	**Exemplo**			
	Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
Overly Literal / Lexical Selection / Ambiguous Translation / Wrong Preposition	W22	*... we'll be able to raise a payment request to our Finance team for you.*	*... poderemos [levantar] um pedido de pagamento [para] a nossa equipe de Finanças para [você].*	levantar/para/ você

Tabela 30 – Exemplos de categorização: *Overly Literal*/*Lexical Selection*/*Ambiguous Translation*/*Wrong Preposition*

Em síntese, considerando a análise apresentada nos parágrafos anteriores, é possível afirmar que entre os 66 dados válidos anotados originalmente sob a etiqueta *Word Order*, 48 revelavam problemas de ordem de palavras: 8 deles envolviam somente a etiqueta *Word Order*; 10 envolviam erros de *Word Order* juntamente com outra(s) etiqueta(s); e 30 dados envolvem problemas mais difíceis de identificar em PB e que, por essa razão, serão tratados somente na seção 6.2.

DADOS ENVOLVENDO FENÔMENOS DE ORDEM DE PALAVRAS	
Erros somente de *Word Order*	8
Erros de *Word Order* com outra(s) etiqueta(s)	10
Dados envolvendo "Possivelmente não natural" e "Variação em PB"	30
TOTAL	**48**

Tabela 31 – Dados válidos envolvendo fenômenos de ordem de palavras

6.1.2.3 Síntese dos erros de categorização

Tendo em conta a análise levada a cabo nos parágrafos anteriores, foi possível encontrar alguns casos em que a categorização foi feita de maneira inadequada pelos anotadores de PB na Unbabel, confirmando uma tendência já observada por Testa (2018) e Figueira (2018) acerca da anotação não uniforme dos dados na ferramenta de anotação dessa empresa. Considerando a anotação feita manualmente no âmbito desta pesquisa, foram mal categorizados 16 entre os 109 dados válidos anotados originalmente em *Agreement*: em 11 não havia nenhum erro e em 5 havia erros envolvendo outras etiquetas. No caso de *Word Order*, dos 66 dados válidos categorizados pelos anotadores, 18 dados foram mal categorizados: 12 não apresentaram nenhum erro e 6 continham erros que envolvem outras etiquetas.

DADOS MAL CATEGORIZADOS	***Agreement***	***Word Order***
Não há erro	11	12
Erros envolvendo outras etiquetas	5	6
TOTAL	**16**	**18**

Tabela 32 – Dados mal categorizados (*Agreement* e *Word Order*)

O foco principal dessa seção é fornecer uma análise sucinta da categorização feita pelos anotadores, ressaltando as principais erros de relativamente ao processo de anotação. A análise apresentada na presente subseção fornece material que poderá auxiliar a empresa na compreensão de alguns dos erros de categorização e no aprimoramento das orientações dadas aos anotadores. Durante esse processo, foi possível fazer retirar os dados que não envolviam erros de concordância e ordem de palavras, fornecendo assim uma lista de dados com fenômenos relevantes para a análise da próxima seção 6.2. Essa análise também auxiliará no desenvolvimento das sugestões e contributos da presente pesquisa, apresentados na seção 7.

6.2 Comentários dos dados

A presente seção busca observar estritamente os dados que contêm erros linguísticos envolvendo concordância ou ordem de palavras, ou que, mesmo não contendo erro, fazem parte de estruturas interessantes para a discussão dessas questões em PB. Assim,

a análise feita na presente seção inclui somente dados relativos a concordância e a ordem de palavras, excluindo-se erros de outras naturezas apresentados em 6.1.

Tendo em vista esse objetivo, foi necessário passar por duas etapas adicionais de filtragem de dados feitas manualmente e denominadas "Separação" e "Adição", que podem ser observadas mais detalhadamente no Anexo 6[33]. Considerando-se a segmentação não uniforme feita em alguns dados pelos anotadores da Unbabel, já discutida na seção 6.1.1, a etapa de "Separação" permitiu uma maior uniformização na segmentação dos dados. Além disso, tendo em vista a análise que será feita na presente seção 6.2, essa etapa permitiu uma melhor visualização e contabilização das diferentes categorias envolvidas nos erros concordância e ordem de palavras encontrados nos dados. Através da etapa de "Adição", foi possível aumentar a diversidade dos dados problemáticos, pois a análise não se limitou às expressões selecionadas pelos anotadores da Unbabel, mas observou todos os textos de chegada fornecidos pela empresa.

A partir desses dois processos de recolha manual, foram adicionados 17 dados pelo processo de Separação e 22 dados pela Adição (tabela 33), no caso da concordância; bem como 7 dados pela Separação e 20 dados através da Adição (tabela 34), no caso de ordem de palavras. Como é possível verificar nas tabelas a seguir, o número de dados total analisados na presente subseção é encontrado através da soma dos dados inseridos manualmente por meio da Adição e da Separação com o número de dados válidos envolvendo fenômenos de concordância e ordem de palavras", já mencionados nas tabelas 21 e 31 da seção 6.1.2.2. Assim, chegou-se a um total de 132 dados problemáticos envolvendo concordância e 75 dados com problemas na ordem de palavras.

DADOS PROBLEMÁTICOS ENVOLVENDO *CONCORDÂNCIA* CONSIDERADOS NA ANÁLISE GRAMATICAL		
Problemas de concordância selecionados pelos anotadores da Unbabel		93
Problemas de concordância inseridos manualmente	***Adição***	22
	Separação	17
TOTAL		**132**

Tabela 33 – Dados problemáticos envolvendo *concordância* considerados na análise gramatical

[33] Os processos de segmentação e seleção dos dados feitos através dessas duas etapas não segue as mesmas regras de segmentação sugeridas pelas *Guidelines*, já discutidas nas seções 4.2.1.2 e 6.1.1, pois serve estritamente o propósito da análise a ser desenvolvida no presente relatório.

DADOS PROBLEMÁTICOS ENVOLVENDO ORDEM DE PALAVRAS CONSIDERADOS NA ANÁLISE GRAMATICAL		
Problemas de ordem de palavras selecionados pelos anotadores da Unbabel		48
Problemas de ordem de palavras inseridos manualmente	***Adição***	20
	Separação	7
TOTAL		**75**

Tabela 34 – Dados problemáticos envolvendo *ordem de palavras* considerados na análise gramatical

À luz do que foi discutido no capítulo 5, em que foi apresentada uma descrição de questões de concordância e de ordem de palavras em PB, as próximas subseções visam fornecer uma análise de alguns desses 132 casos com fenômenos de concordância e 75 casos problemáticos[34] envolvendo a ordem de palavras sob uma perspectiva linguística, procurando apresentar explicações acerca das categorias gramaticais mais envolvidas nos erros e problemas encontrados e das possíveis motivações atrás de tais casos. Também serão sublinhadas algumas características inerentes à língua portuguesa em geral e ao PB mais especificamente que dificultam o processo de tradução e o estabelecimento de um padrão único de concordância e de ordem de palavras. A presente subseção possui a seguinte organização: na seção 6.2.1, são discutidos os dados envolvendo concordância; em 6.2.2, são observados os dados envolvendo ordem de palavras.

6.2.1 Dados envolvendo concordância

Os problemas de concordância encontrados nos dados foram divididos em três grupos: Internos ao SN, Envolvendo Sujeito e Outras estruturas. Foram encontrados 97 dados com erros envolvendo a partilha de traços entre o núcleo do SN e/ou especificadores, modificadores, complementos adjetivais. Como é possível observar na tabela 35, 73% dos erros de concordância encontrados nos dados se manifestava internamente ao SN. Foram encontrados 30 casos de problemas de concordância envolvendo sujeito. Esse tipo de problema foi subdividido em três partes: as relações de concordância entre Sujeito-Verbo

[34] No caso da ordem de palavras, preferiu-se não utilizar a denominação "erro" visto que alguns casos são efetivamente erros, mas outros são posições menos naturais ou que dão origem a leituras diferentes das do texto de partida. Esses casos serão vistos mais adiante na seção 6.2.4. Também na concordância preferiu-se não usar o termo "erro", pois há casos que envolvem o fenômeno de variação entre os traços singular/plural ou concordância semântica, muito presente em PB (cf. seção 5).

(12 casos), entre Sujeito-Predicativo (12 casos) e entre Sujeito-Particípio Passivo (4 casos). O grupo "Outras estruturas" inclui casos diversos em que se verificam também relações de concordância em PB: o pronome relativo, o predicativo do objeto direto e os clíticos.

DADOS ENVOLVENDO *CONCORDÂNCIA*	
Internos ao SN	93
Envolvendo Sujeito	28
Outras estruturas	11
TOTAL	**132**

Tabela 35 – Dados envolvendo *Concordância*

Os próximos parágrafos se dedicam à análise dessas relações, com o objetivo de compreender a origem e complexidade de certos erros. Dado que se pretende explorar as razões que podem estar na origem dos erros e que os textos de partida estão escritos em inglês, apresenta-se em 6.2.1.1 uma descrição breve do inglês relativamente à algumas questões de concordância. As subseções seguintes foram separadas segundo a divisão apresentada na tabela 35. A subseção 6.2.1.2 trata dos dados com problemas de concordância internamente ao SN; 6.2.1.3 trata dos casos de concordância com o sujeito; 6.2.1.4 aborda as outras estruturas com problemas de concordância encontradas nos dados. Adicionalmente, será abordada a relação entre a concordância e a coesão referencial em 6.2.1.5 e uma síntese geral relacionando os fênomenos tratados e os traços de concordância será fornecida em 6.2.1.6.

Quando se considerou útil para a análise, tabelas com os números de erros segundo as categorias gramaticais foram fornecidas. Nesta seção, por questões de espaço, optou-se por manter somente as partes dos textos de partida e de chegada interessantes para a análise, deixando-se de lado informações como os segmentos anotados e as categorias selecionadas. Os exemplos apresentados entre as seções 6.2.1.2 e 6.2.1.5 foram retirados dos dados fornecidos pela empresa. Exceto se expressamente indicado, os exemplos dessas seções seguem a seguinte organização: a alínea (a) apresenta o texto na LP; a alínea (b), contendo o erro ou o fenômeno de concordância que será discutido, apresenta a tradução na LC realizada pela empresa; as alíneas (c) e (d) são propostas de tradução ou exemplos de tradução que podem auxiliar a discussão feita.

6.2.1.1 Concordância em inglês

As relações de concordância em PB foram discutidas na seção 5. No entanto, para compreender a origem de certos erros, será necessário observar sucintamente algumas das características do funcionamento da concordância em gênero, número e pessoa da língua inglesa que mais a diferenciam do PB e podem dificultar o processo de tradução. Os parágrafos seguintes se dedicam a fornecer informações que possibilitem a comparação do funcionamento da concordância entre inglês e PB. Será dada atenção especial às características da concordância dentro do SN, tendo em vista o número de erros desse tipo encontrados nos dados.

O SN em inglês é composto por um núcleo, que pode ser introduzido por um ou mais determinantes e pode apresentar modificadores antes ou depois desse núcleo (Greenbaum, 1996: 205). Diferentes categorias gramaticais e unidades sintáticas podem modificar o núcleo do SN. A lista (197) apresentada abaixo ilustra o uso de alguns desses modificadores: adjetival (197a), participial reconvertido em adjetival (197b), gerundivo reconvertido em adjetival (197c), nominal (197d), preposicional (197e) e oracional (oração relativa, 226f).

(197)

a. *A **great** book*
b. *Many **broken** dishes*
c. *A **confusing** movie*
d. *An **office** chair*
e. *Twelve mice **without tails***
f. *The car **that was in the garage***

Quanto aos complementos do SN em inglês, como em PB, eles são licenciados pelo núcleo do SN que precisa dessa informação suplementar para completar o seu sentido. Segundo Payne (2010: 3), os complementos sempre seguem o núcleo do SN. Esses elementos são tipicamente SP ou orações, como é possível observer nos exemplos (198a) e (198b), respectivamente.

(198)

a. *a belief **in the hereafter***
b. *the rumor **that she was running for office***

(Payne, 2010: 9)

No caso dos determinantes, em inglês, não há marcação de gênero no artigo definido "*the*" e nos artigos indefinidos "*a*" e "*an*". O artigo definido não flexiona, mas pode determinar nomes no singular ou plural, como no exemplo (199a). Já os artigos indefinidos somente podem ser utilizados com nomes contáveis no singular (Greenbaum, 1996: 164), como no exemplo (199b).

(199)

a. ***the*** *issue* / ***the*** *issues*
b. ***an*** *issue* / ****an*** *issues* / ***some*** *issues*

(Greenbaum, 1996: 164)

Os demonstrativos "*that*", "*this*", se flexionam em número, sendo seus correspondentes plurais as expressões "*those*" e "*these*", respectivamente. Entretanto, não têm marcação explícita de gênero. Como apresentado nas expressões apresentadas em (200a) e (200b) os dois primeiros demonstrativos, "*that*" e "*this*", manifestam o singular e os dois últimos, "*those*" e "*these*", se encontram no plural.

(200)

a. ***this*** *girl* / ***this*** *boy* / ***that*** *girl* / ***that*** *boy*
b. ***those*** *girls* / ***those*** *boys* / ***these*** *girls* / ***these*** *boys*

Segundo Greenbaum (1996: 165), os três principais conjuntos de pronomes em inglês, nomeadamente o pessoal, o possessivo e o reflexivo, estão correlacionados. Eles exibem contrastes em pessoa, número e gênero. Porém, esse autor destaca que tais contrastes não estão disponíveis em todas as instâncias. Para exemplificar, o contraste de gênero masculino/feminino somente é expresso na 3ª pessoa desses três conjuntos, como em (201a) e (201b). Alguns pronomes, como os relativos, podem expressar ambos os gêneros, como os apresentados em (202a), e outros são considerados neutros e não expressam nenhum gênero, como em (202b).

(201)

a. masculino: *he*, *him*, *his*, *himself*
b. feminino: *she*, *her*, *hers*, *herself*

(202)

a. masculino ou feminino: *who*, *whom*, *whoever*, *whomever*
b. impessoal: *it*, *its*, *itself*, *which*

(Greenbaum, 1996: 107)

Por sua vez, os possessivos com função de determinante em inglês mudam segundo as três pessoas gramaticais do possuidor (203), mas, diferente do PB, não flexionam em número de acordo com a entidade possuída (204a). É de notar que, na 2ª pessoa, os pronomes pessoais e possessivos possuem a mesma forma para o singular e para o plural (203b).

(203)

a. 1ª pessoa*: my*, *our*
b. 2ª pessoa: *your*
c. 3ª pessoa: *his*/*her*, *their*

(Greenbaum, 1996: 166)

(204)

a. *...all **my** papers are in a mess in **my** desk*

(Greenbaum, 1996: 99)

O gênero não é morfologicamente marcado em adjetivos nem em particípios e gerúndios com valor adjetival em inglês. Logo, os elementos dessas categorias gramaticais não sofrem alterações de acordo com o gênero do nome que eles modificam e também não expressam número na sua forma. Os excertos abaixo exemplificam essa ausência de marcas no adjetivo (205a), no particípio (205b) e gerúndio (205c) reconvertidos em adjetivais:

(205)

a. *The **tall** girl* / *The **tall*** boy / *The **tall** children*
b. *The **excited** girl* / *The **excited** boy* / *The **excited** children*
c. *The **boring** girl* / *The **boring** boy / The **boring** children*

Os nomes comuns podem ser divididos em contáveis ou não-contáveis. Esses últimos possuem um sentido de coletividade e se manifestam sempre no singular (Greenbaum, 1996: 97-98), como nos exemplos (206a) e (206b) retirados do texto do autor. Os contáveis possuem formas singular e plural e podem ser introduzidos por determinantes ou por quantificadores que fazem distinções de número como exemplifica o nome singular em (207a) e plural em (207b).

(206)

a. *the/this/some/any/no information*
b. *my/whose/which/what/whatever sugar*

(207)

a. *two/several/few/many/these pictures*
b. *one/every/either/this picture*

(Greenbaum, 1996: 97-98)

Em resumo, atualmente o inglês não possui gênero gramatical: os nomes não diferenciam o gênero através de marcas flexionais e os determinantes e adjetivos não variam segundo o gênero dos nomes que acompanham ou modificam.

Como já foi mencionado, os pronomes pessoais de primeira e terceira pessoa possuem contrastes em número singular e plural: *I* (1ªsg) e *we* (1ªpl); *he*/*she*/*it* (3ªsg) e *they* (3ªpl) (Greenbaum, 1996: 166). Já o pronome pessoal de segunda pessoa é idêntico no singular e no plural: *you* (2ªsg e 2ªpl). Apesar de se manifestar nessas três pessoas gramaticais, quanto à conjugação verbal, o inglês se diferencia muito do português, pois somente a terceira pessoa do singular do tempo presente é morfologicamente marcada, com o sufixo "*-s*". Com exceção do verbo "*be*", as formas verbais em outras pessoas gramaticais são idênticas, como exemplificam (208a) e (208b) (Greenbaum, 1996: 237).

(208)

a. *I/You/They/We* ***eat*** salad regularly.
b. *He/She* ***eats*** salad regularly.

Quanto à concordância verbal, o verbo em inglês também deve concordar em pessoa e número com o núcleo do sujeito. Dado que o sistema flexional nos verbos é pobre em inglês, a ocorrência de sujeito nulo está excluída desta língua, ao contrário do que acontece em PB, que ainda mantém sujeitos nulos em certos (mas não em todos os) contextos (ver seção 5). Isso fica explícito quando se compara os excertos (209b) e (210b):

(209)

a. *I eat salad regularly.*
b. **Eat salad regularly.*

(210)

a. *Eu como salada.*
b. [-] Como salada.

Como aponta Greenbaum (1996: 237), se o sujeito possuir um nome como núcleo do SN, a regra geral é que o verbo no tempo presente possui o sufixo "*-s*" quando o nome é singular (211a) e não possui esse sufixo quando no plural (211b).

(211)

a. *His account **contains*** many historical solecisms.
b. *Many terrestrial soils*, in contrast, ***contain*** large proportions of very small particles made up of clay minerals.

(Greenbaum, 1996: 237)

Vale fazer um paralelo entre a concordância de nomes coletivos (*collective nouns*) e de plurais coletivos (*collective plurals*) em inglês. Segundo Payne (2010: 120-121), os primeiros podem provocar concordância plural, quando os indivíduos da coletividade atuam separadamente, ou singular, quando a coletividade age como um só organismo. Os excertos em (212) exemplificam o uso do nome coletivo "*government*" com o verbo singular (212a) e plural (212b). Já os plurais coletivos somente desencadeiam a concordância plural, como indicam as agramaticalidades em (213).

(212)

a. *However, the Government **has** no intention to privatize health care.*
b. *I'm ashamed os some of the things the government **have** done.*

(Payne, 2010: 121)

(213)

a. *the people **are**/*__**is**__ overemployed*
b. *Cattle **have**/*__**has**__ very sensitive muzzles.*

(Payne, 2010: 120)

6.2.1.2 Concordância interna ao SN

Foram encontrados 69 erros envolvendo a concordância dos especificadores com o nome, como é possível verificar na tabela 36. A categoria gramatical mais envolvida nos erros de concordância é a dos artigos definidos, correspondendo a aproximadamente 55% do total. A segunda categoria gramatical mais encontrada é a dos possessivos: 15 dados com concordância inadequada continham esse tipo de elemento. Os possessivos foram inseridos nesta tabela tendo em vista a discussão sobre o funcionamento deles enquanto especificadores em PB, já explicitado na seção 5. Encontrou-se também 10 erros com artigos indefinidos, 2 com demonstrativos e 4 com quantificadores.

CONCORDÂNCIA: ESPECIFICADORES		
Determinantes	***Artigo Definido***	38
	Artigo Indefinido	10
	Demonstrativo	2
Possessivos		15
Quantificadores		4
TOTAL		**69**

Tabela 36 – Concordância: Especificadores-Nome

Quanto às estruturas mais típicas com artigos definidos, foram encontrados erros em que esse determinante era o único elemento antes do núcleo do SN, como em (214b), ou acompanhado de possessivo, ambos mal flexionados, como em (215b). Note-se que enquanto no exemplo (214a) o artigo "*the*" está explicitamente escrito na LP, no exemplo (215a) somente está no texto de chegada o possessivo "*your*", como é característico do inglês. Nesse último caso, o determinante foi então adicionado durante o processo de tradução para o PB.

(214)

a. *you forgot **the** password.*
b. *você esqueceu **o** senha.
c. você esqueceu **a** senha

(215)

a. *on your account*
b. *n**o** seu conta
c. n**a** sua conta

Também foram encontrados artigos definidos que determinavam nomes com a forma da LP (isto é, estrangeirismos), como em (216b). A tradução nesses casos é desafiadora, pois é necessário conhecer os traços do termo estrangeiro na LC para que possa ser feita a concordância adequada. No exemplo dado, o termo "*App*" corresponde ao termo "aplicativo" em PB. Em (216b) houve, provavelmente, uma interferência do PE dado que nesta variedade do português a tradução de "*App*" é "aplicação", usando-se assim o artigo definido feminino mesmo nos casos em que a palavra é truncada. Como é possível

observar em (216c), a tradução mais adequada para o artigo definido seria "**o** *App*" ou "**o** aplicativo", caso se opte por traduzir o termo em PB.

(216)

a. *in (COMPANY M) App*
b. * n**a** App (COMPANY M)
c. n**o** aplicativo (COMPANY M) / n**o** *App* (COMPANY M)

Outro caso interessante é a presença do símbolo "/" em certas palavras, como exemplifica o nome "*purchase/s*" em (217a). Esse símbolo pode ser utilizado em inglês para representar a alternância de traços entre o singular e o plural. Em PB, o símbolo mais adequado para marcar essa alternância seriam os parênteses "()". Tendo em vista a necessidade de marcar a concordância também nos determinantes, a tradução mais adequada seria a expressão apresentada em (217c). Esse é um dos casos inseridos na categoria de "Flexão de Número", citados na tabela 37, apresentada mais à frente.

(217)

a. *purchase/s*
b. *a compra **/ s**
c. a**(s)** compra**(s)**

Foram encontrados dados em que artigo indefinido era o único elemento à esquerda do nome, como em (218b). Em outros dados, essa categoria gramatical estava seguida de adjetivo, ambos em posição pré-nominal, como em (219b).

(218)

a. *a travel search engine*
b. ***um** ferramenta de busca de viagens
c. **uma** ferramenta de busca de viagens

(219)

a. ***a** new card*
b. ***uma** nova cartão
c. **um** novo cartão

Também foram encontrados casos envolvendo estrangeirismo (220) e o símbolo "/" nos dados com erros de artigo indefinido (221). No caso de (221a), a alternância não é entre os traços singular e plural, mas entre dois elementos inseridos no SN. Nesse caso, o método mais adequada de traduzir seria a inserção da conjunção "*ou*" e a repetição do

artigo indefinido nos dois termos, estabelecendo-se a necessária relação de concordância, como em (221c).

(220)

a. ***a*** *Venture Builder*
b. *__um__ Venture Builder
c. **uma** *Venture Builder*

(221)

a. ***an*** *airline/travel agent*
b. *__um__ linha aérea/agente de viagens
c. **uma** linha aérea ou um agente de viagens

Os dois casos de erro com demonstrativo envolviam a concordância entre os traços de gênero, como o exemplo apresentado em (222):

(222)

a. ***this*** *account*
b. *__este__ conta
c. **esta** conta

Os possessivos exercem a função de especificador em inglês e, nesta língua, não são precedidos de artigo, como em (223a) e (224a). Durante a tradução, os possessivos foram traduzidos sem artigos, como é demonstrado em (223b), ou com um artigo definido como em (224b). A questão da obrigatoriedade no uso do artigo juntamente com o possessivo, já citada na seção 5, não foi considerada na presente seção tendo em vista as especificidades do PB e os objetivos desta pesquisa.

(223)

a. ***your*** *booking confirmation*
b. *__seu__ confirmação de reserva
c. **sua** confirmação de reserva

(224)

a. ***your*** *workouts*
b. *a **sua** treinos
c. os **seus** treinos

Foram encontrados erros com quantificadores universais e vagos, como em (225b) e (226b), respectivamente. Nos dois exemplos há erro envolvendo o traço de número e os

quantificadores co-ocorrem com um artigo definido e um adjetivo (225b) ou com um determinante indefinido (226b).

(225)

a. ***all*** *the great content*
b. ***todos** o grande conteúdo
c. **todo** o grande conteúdo

(226)

a. ***Any*** *other email adresses*
b. ***Qualquer** outro endereços de e-mail
c. **Quaisquer** outros endereços de e-mail

Apesar de serem facilmente identificáveis pelos anotadores e editores, os erros envolvendo a concordância entre os especificadores e o nome são compreensíveis tendo em vista as diferenças entre o inglês e o PB. A tabela 37 a seguir apresenta uma síntese dos erros encontrados segundo os tipos de concordância: 43 erros somente de gênero; 13 somente de número; e 9 de gênero e número ao mesmo tempo. Nos dados envolvendo especificadores, as expressões "Palavra Estrangeira" e "Flexão de Número" designam, respectivamente, casos especiais em que o erro inclui estrangeirismos (**um** *Venture Builder*, 249b) e em que a alternância de traços não foi feita de maneira adequada (a compra **/ s**, 246b).

TIPOS DE CONCORDÂNCIA : ESPECIFICADORES		
Concordância de	***Gênero***	43
	Número	13
	Gênero e Número	9
Palavra Estrangeira		3
Flexão de Número		1
TOTAL		**69**

Tabela 37 – Tipos de concordância: Especificadores

Quando se observa mais detalhadamente a relação entre as categorias gramaticais dos especificadores e os traços de concordância nos dados analisados, é possível verificar que o gênero é o traço com o maior número de erros nos dados com artigos definidos, artigos indefinidos e possessivos. Os dados com quantificadores possuem somente erros relacionados ao número. Também é interessante notar que os demonstrativos e os

possessivos encontrados nos dados fornecidos pela empresa não apresentaram erro de número. O artigo definido e o possessivo foram as únicas categorias de especificadores em que se encontrou simultaneamente erro de gênero e número.

TRAÇOS DE CONCORDÂNCIA E CATEGORIAS GRAMATICAIS: ESPECIFICADORES			
Categoria Gramatical	**Gênero**	**Número**	**Gênero e Número**
Artigo Definido	21	8	6
Artigo Indefinido	8	1	0
Demonstrativo	2	0	0
Possessivo	12	0	3
Quantificadores	0	4	0
TOTAL	43	13	9

Tabela 38 – Traços de concordância e categorias gramaticais: Especificadores

Foram encontrados 5 dados em que o núcleo do SN manifestava problemas de concordância, fidelidade ou flexão, como apresentado na tabela 38. Os problemas envolvendo esse tipo de elemento foram separados em duas categorias: "Concordância de Numeral" e "Flexão de Número". No primeiro caso, há efetivamente falta de concordância entre os traços do numeral e do núcleo do SN, como em "3 **mês**" (227b). No segundo caso, há falta de fidelidade aos traços do núcleo no texto de partida ((228a) *maximum* ***durations***; (228b) **duração** máxima); ou problemas na sua grafia ((229b) (5) **Imagem**).

TIPOS DE CONCORDÂNCIA : NÚCLEO	
Concordância de Numeral	3
Flexão de Número	2
TOTAL	**5**

Tabela 39 – Tipos de concordância: Núcleo

Um exemplo de erro de "Concordância de Numeral" é apresentado no excerto (227b): o núcleo "*mês*" não concorda com o numeral plural "*3*" que o antecede, sendo correta a proposta de tradução (227c).

(227)

a. *3* ***month***
b. *3 **mês**
c. 3 **meses**

Já em (228), há falta de fidelidade ao texto de origem, pois o núcleo "*durations*" está no plural em inglês (228a), mas foi traduzido no singular (228b). Esse tipo de erro foi assinalado como contendo erro de "Flexão de Número". O dado (228) é bastante interessante tendo em vista o envolvimento de outras classes de palavras e da ordem dessas palavras no erro, por isso ele será revisto mais adiante.

(228)

a. *workout minimum/maximum* ***durations***
b. *treino mínimo / **duração** máxima
c. **durações** mínimas e máximas de treino

No caso de (229b), o nome "*Screenshot*" foi traduzido por uma longa expressão nominal em PB, causando erro na inserção da marca de plural "(s)". Em PB, a palavra "Imagem" é o núcleo do SN, logo a marcação de plural deveria estar após esse nome para representar a alternância do traço de número. Apesar de também envolver um numeral, como em (229a), esse caso não foi inserido como erro de concordância de número nos dados, mas sim como erro de "Flexão de Número", pois a tradução pela forma plural "Imagens" não resolveria o problema de tradução: o que falta em (229b) é a citada alternância, proposta em (229c).

(229)

a. *(5)* ***Screenshot(s)***
b. (5) Imagem De Captura De **Tela(s)**
c. (5) Imagem(ns) de captura de **tela**

Considerem-se, agora, os modificadores (incluindo os derivados de particípios e de gerúndios na LP e/ou na LC), que devem concordar também em gênero e número com o núcleo do SN. Tendo em vista as diferenças entre o inglês e o PB quanto aos elementos passíveis de serem inseridos no SN, considerou-se relevante fazer um paralelo entre as categorias gramaticais dos modificadores na LP e na LC (tabela 40). É possível observar que nos casos em que o modificador em inglês é um gerúndio convertido em adjetivo ou um nome, esses elementos foram traduzidos por adjetivos nos dois casos encontrados.

Os erros de concordância com modificadores que envolviam adjetivos na LP traduzidos por adjetivos correspondentes na LC correspondem a aproximadamente 68% dos casos. Considerando-se as estruturas do SN no texto de partida, foram encontrados 13 erros com adjetivos, 4 erros envolvendo formas participiais, 1 erro envolvendo a forma gerundiva e 1 erro envolvendo o nome.

CONCORDÂNCIA : MODIFICADORES		
No texto de partida	**No texto de chegada**	
Adjetivo	***Adjetivo***	13
Forma participial	***Forma participial***	4
Forma gerundiva	***Adjetivo***	1
Nome	***Adjetivo***	1
TOTAL		**19**

Tabela 40 – Estruturas de concordância: Modificadores

Nos exemplos apresentados abaixo, os adjetivos traduzidos contêm erros de concordância de número (230b) e de gênero (231b). Em (230a) o adjetivo "*visual*" modifica o nome "*issues*", não se mantendo esta relação na tradução feita em (230b). Já em (231b), o núcleo do SN em inglês, "*currency*", foi traduzido pela expressão "moeda corrente", modificando o adjetivo "estrangeiro" todo esse grupo e devendo concordar com o seu núcleo "moeda", como em (231c).

(230)

a. *Most **visual** issues*
b. *A maioria dos problemas **visual***
c. A maioria dos problemas **visuais**

(231)

a. *a **foreign** currency*
b. *uma moeda corrente **estrangeiro***
c. uma moeda corrente **estrangeira**

O exemplo (232), abaixo, já apresentado acima (228), contitui um caso interessante, pois envolve o símbolo "/". Na expressão em inglês, esse símbolo representa a alternância entre os dois adjetivos "*minimum/maximum*" que modificam ambos o núcleo

"*durations*". Junta-se a esses elementos o nome "*workout*" que é núcleo de um SN complemento de *durations*. No entanto, na tradução feita em (232b), esse símbolo representa uma alternância entre dois grupos nominais diferentes: "treino mínimo" e "duração máxima". O problema é que, na LP, "mínimo" não modifica "treino", mas sim "durações". Além disso, "treino" faz parte do complemento do núcleo "durações" e em PB deveria estar inserido no SP "de treino", como sugerido em (232c). Dado que os adjetivos em inglês não exibem traços visíveis de número, a identificação das relações de concordância pode ser desafiadora durante a tradução. Contudo, neste caso, a tradução "treino mínimo" deveria ser excluída imediatamente, considerando-se que em inglês os adjetivos se posicionam tipicamente antes do nome que modificam e que neste caso os adjetivos "*minimum/maximum*" precedem apenas o nome "*durations"*. Esse exemplo mostra bem a importância de se identificar corretamente as relações de concordância entre o nome e os outros constituintes do SN e, para além disso, mostra como a ocorrência de vários erros diferentes na mesma estrutura pode dificultar a anotação e, consequentemente, a correção.

(232)

a. *workout* ***minimum/maximum*** *durations*
b. *treino **mínimo** / duração **máxima***
c. durações **mínimas/máximas** de treino

Relativamente a formas participiais convertidas em adjetivos, há casos, como (233b), em que a agramaticalidade é flagrante por existir somente um nome no SN e o particípio não partilhar os traços desse nome (para além da ordem, que será objeto de análise mais adiante):

(233)

a. ***pre-paid*** *financial institutions*
b. ***pré-pago** instituições financeiras*
c. instituições financeiras **pré-pagas**

Casos como o apresentado em (234) foram inseridos como erros internamente ao SN. Em (234b), o adjetivo deve concordar com o nome do destinatário da mensagem em PB. A palavra foi anonimizada manualmente, mas tal anonimização manteve uma referência ao traço de gênero (feminino) e ao número (singular), logo a tradução apresentada em (234c) é a única possível.

(234)

a. *Dear (PERSON'S NAME F),*
b. ***Querido** (PERSON'S NAME F),
c. **Querida** (PERSON'S NAME F),

O exemplo (235) ilustra outro caso em que é difícil identificar de maneira exata quais são as relações de dependência que se estabelecem internamente ao SN. Em (235a) "*registered*" pode ser modificador de "*email*" ou de "*email inbox*", registando-se, portanto, uma ambiguidade estrutural. Na tradução feita pelos anotadores da Unbabel, em (235b), esse particípio modifica "caixa de entrada de e-mail". Na proposta de tradução (235c), esse elemento está inserido no SN que é constituinte do SP (do e-mail registrado) e modifica "e-mail". Apesar de não existir erro gramatical em (235b), tendo em vista o contexto discursivo, a tradução mais adequada em PB seria a proposta apresentada em (235c), pois de maneira geral se registra e-mails e não caixas de entrada.

(235)

a. *your **registered** email inbox*
b. sua caixa de entrada de e-mail **registrada**
c. sua caixa de entrada do e-mail **registrado**

A tradução de formas gerundivas do inglês com função de modificador também se revela desafiadora, dado que expressões dessa categoria não podem funcionar como modificadores do SN em PB. Por isso, o tradutor deve optar por outras categorias gramaticais durante a tradução. No caso de (236b), o gerúndio foi traduzido por um adjetivo, mas houve erro de concordância de número.

(236)

a. *the **following** details*
b. *o **seguinte** detalhes
c. os **seguintes** detalhes

Diferente do inglês, em PB um SN não pode exercer função de modificador de outro grupo nominal, pelo que a tradução tem de passar pela alteração da categoria, mais especificamente, deverá ser traduzido por um SP ou por um SA. O SN "*bank*" em (237a) pode ser modificador de "*payment*" ou de "*payment restrictions*". A tradução encontrada nos dados (237b) é agramatical, pois os traços de "bancários" (masculino e plural) não são completamente compatíveis nem com os traços de "pagamento" (masculino e singular) nem com os traços de "restrições de pagamento" (feminino e plural). A opção (237c)

ilustra o caso em que o elemento "bancários" funciona como modificador de "restrições de pagamento" e a opção (237d) ilustra esse adjetivo como modificador de "pagamento". Caso a opção (237c) seja considerada a mais adequada, houve erro de gênero na tradução feita pela empresa, apresentada em (237b). Caso a opção (237d) seja considerada correta, houve erro de número em (237b). Tendo em vista as várias possibilidades de tradução, esse dado foi contabilizado como erro de "Concordância de Número ou Gênero" nos dados da presente pesquisa.

(237)

a. *any bank or international payment restrictions*
b. *quaisquer restrições de pagamento **bancários** ou **internacionais**
c. quaisquer restrições de pagamento **bancárias** ou **internacionais**
d. quaisquer restrições de pagamento **bancário** ou **internacional**

O erro de concordância envolvendo a tradução de "*international*" em (237a) também foi inserido na análise enquanto problema no adjetivo. Nesse caso, o adjetivo "internacional" em PB somente exibe visivelmente o traço de número, sendo a forma do masculino igual à do feminino (adjetivo uniforme). Por isso, ao fazer o paralelo entre as opções apresentadas em (237c) e (237d) e a tradução feita na empresa (237b), é possível encontrar duas possibilidades: se a forma "internacionais" se refere às "restrições de pagamento", não há erro de tradução em (237b); mas se esse elemento modifica somente "pagamento" há erro de concordância de número em (237b). Por isso, esse termo foi assinalado como contendo erro de "Concordância de Número ou Não há erro".

Para resumir, na tabela 41 seguinte é feito um paralelo entre as categorias gramaticais na LP e os traços envolvidos nos dados de concordância. Dentre os modificadores com problemas de concordância, 9 contêm erros somente de gênero (uma moeda corrente **estrangeiro**, 231b); 6 contêm erros somente de número (o **seguinte** detalhes, 235b); e 2 elementos contêm erro de gênero e número ao mesmo tempo (**pré-pago** instituições financeiras, 232c). A maior parte dos erros com adjetivos e particípios convertidos em adjetivos envolvia somente o traço de gênero: 7 e 2 casos, respectivamente. O único dado com o modificador gerúndio convertido em adjetivo envolve o traço de número. Os casos de "Gênero *ou* Número" e "Número *ou* Não há erro" se referem aos erros já citados em (237b) "quaisquer restrições de pagamento **bancários** ou **internacionais**" em que há duas possibilidades de modificação para o nome "*bank*" e para o adjetivo "*international*". Por isso, eles foram as únicas categorias gramaticais contabilizadas nessas duas colunas.

TRAÇOS DE CONCORDÂNCIA E CATEGORIAS GRAMATICAIS: MODIFICADORES					
Categoria gramatical na LP	**Traços envolvidos**				
	Gênero	***Número***	***Gênero e Número***	***Gênero ou Número***	***Número ou Não há erro***
Adjetivo	7	4	1	0	1
Forma Participial	2	1	1	0	0
Forma Gerundiva	0	1	0	0	0
Nome	0	0	0	1	0
TOTAL	9	6	2	1	1

Tabela 41 – Traços de concordância e categorias gramaticais: Modificadores

6.2.1.3 Concordância com o sujeito

Os dados problemáticos envolvendo concordância entre o sujeito e o verbo totalizaram 12 casos. Como já foi tratado na seção 5, a concordância verbal no PB terá em consideração fênomenos de variação e a concordância por sentido. Por isso, não é possível afirmar de maneira generalizada que todos esses dados contêm erro de concordância. Essa questão será discutida brevemente através dos exemplos apresentados nos parágrafos seguintes. Dado o comportamento diferenciado entre os verbos finitos e no infinitivo flexionado, os dados foram tratados separadamente, como é possível verificar na tabela 42. Vale sublinhar que em ambos os casos, não foram encontrados nos dados erros ou casos problemáticos envolvendo o traço de pessoa.

CONCORDÂNCIA : SUJEITO-VERBO	
Verbo finito	6
Verbo no infinitivo flexionado	6
TOTAL	**12**

Tabela 42 – Concordância: **Sujeito-Verbo**

No caso dos verbos finitos, o excerto (238) exemplifica um erro com verbo copulativo: em (238a) o verbo "*is*" concorda com o núcleo do SN "*Account*". Em (238b) o verbo "*são*"

não partilha os traços de número com o núcleo do sujeito, "*Conta*", por isso a proposta (238c) seria a mais adequada.

(238)

a. *[A (COMPANY M) Account]* ***is*** *for personal use only*
b. *[As Conta de (COMPANY M)] **são** apenas para uso pessoal
c. [A Conta de (COMPANY M)] **é** apenas para uso pessoal

A concordância sujeito-verbo nos excertos seguintes é mais complexa, pois envolve uma coordenação de SVs, pelo que apenas um sujeito é realizado, "*All documents"*. Em (239b), o sujeito da locução verbal "*deve ser*" corresponde ao SN plural [Todos os documentos]. Esse SN plural está longe da locução verbal, mas a concordância de pessoa e número deve ser mantida. Por isso, a proposta apresentada em (239c) é a mais adequada.

(239)

a. *[All documents] must: • Show the issue date and [-]* ***should not be*** *older than 3 months*
b. *[Todos os documentos] devem: • Mostrar a data de emissão e [-] **não deve ser** superior a 3 meses
c. [Todos os documentos] devem: • Mostrar a data de emissão e [-] **não devem ser** superiores a 3 meses

No excerto apresentado em (240), verifica-se aquilo que é comumente designado como concordância semântica. De maneira geral, o verbo deve concordar com o núcleo do SN, que, no caso de (240b), é gramaticalmente singular: *maioria*. Mas também existe a possibilidade de o verbo concordar com o sentido plural desse núcleo, fazendo assim a chamada concordância semântica ou por sentido. A locução verbal "*can be*" em (240a) foi traduzida de duas formas distintas pelos tradutores da empresa: no plural em (240bi) e no singular em (240bii). Tendo em vista a variação que afeta o PB, não é possível determinar de maneira indiscutível qual das duas formas é a mais adequada, por isso na presente análise ambas foram assinaladas com a expressão "Envolvendo concordância semântica".

(240)

a. *[Most visual] issues* ***can be****...*
b.
    i. [A maioria dos problemas visuais] **podem ser**...

ii. [A maioria dos problemas visuais] **pode ser** ...

A expressão "*purchase/s*", em (241), já mencionada na subseção acerca da concordância no interior do SN, foi assinalada como contendo erro de "Flexão de número" devido à ocorrência do símbolo "/". O verbo "*será*" foi utilizado impropriamente em (241b), pois a alternância entre os traços singular e plural, presente no núcleo do SN sujeito, também deveria ter sido inserida no verbo, como na proposta (241c). Este caso foi assinalado com a expressão "Flexão de Número" durante a análise, tendo em vista ser necessária a inserção do símbolo"()" para marcar essa alternência de traços.

(241)

a. *[purchase/s]* ***will be*** *deducted*
b. [a compra / s] **será** deduzida
c. [a(s) compra(s)] **será(ão)** deduzida(s)

Como já foi mencionado na seção 5, o infinitivo flexionado está sujeito a variação em PB, pois em muitos casos o uso da flexão varia de acordo com o contexto. No dado (242b) apresentado a seguir há efetivamente erro de concordância, pois é possível constatar, através dos traços do verbo "Adicione", que o destinatário da mensagem é o interlocutor, pelo que o sujeito nulo deve ser singular e, por isso, o infinitivo não pode ser plural, como apresentado em (242b). Assim, sugere-se a tradução (242c) em que o verbo "ficar" está na 3ª pessoa do singular.

(242)

a. *[-] Add (...) to reduce the likelihood of [-]* ***getting locked out*** *of your account*
b. [-] Adicione (...) para reduzir a probabilidade de [-] **ficarem** bloqueados na sua conta
c. [-] Adicione (...) para reduzir a probabilidade de [-] **ficar** bloqueado na sua conta

É de notar que a flexão do infinitivo na 3ª pessoa do singular não se diferencia da sua forma não-flexionada (como já tratado na seção 5). No caso de um sujeito na 3ª pessoa do plural, como em (243c), a diferença entre a forma flexionada e não flexionada é visível. Na tradução apresentada em (243b), é obrigatório o compartilhamento de traços entre o particípio e sujeito, por isso a agramaticalidade. Se o particípio for corrigido, como em (243c), a falta de flexão no verbo "ser" causa estranhamento, tendo em vista a

proximidade entre esses elementos. Por isso, a tradução mais adequada seria a (243d) em que ambos o verbo e o particípio estão flexionados no plural.

(243)

a. *[some vídeos] are taking some time **to be** subtitled*
b. *[alguns vídeos] levam algum tempo para **ser** legendado
c. ?[alguns vídeos] levam algum tempo para **ser** legendados
d. [alguns vídeos] levam algum tempo para **serem** legendados

Nos excertos apresentados em (244), a referência do sujeito dos infinitivos é fixada pelo objeto direto da frase superior [*travellers*/viajantes]. Aqui o fenômeno de variação no uso do infinitivo fica evidente: as opções com infinitivo não-flexionado (244b) e com infinitivo flexionado (244c) estão ambas corretas em PB. Esse tipo de dado foi assinalado como "Variação em PB" na presente pesquisa e, apesar de não existir erro na tradução feita pela empresa (244b), considerou-se ser um caso interessante para a discussão acerca da concordância em PB.

(244)

a. *a tool to help [travellers] **plan** their trips and **find** the best travel arrangement*
b. uma ferramenta para ajudar [viajantes] a **programar** as suas viagens e **encontrar** o melhor acordo de viagem
c. uma ferramenta para ajudar [viajantes] a **programarem** as suas viagens e **encontrarem** o melhor acordo de viagem

A maioria dos exemplos apresentados nos parágrafos anteriores, relativos à concordância sujeito-verbo, envolvia também a concordância entre o sujeito e seu predicativo. Esses casos também foram contabilizados e observados na presente análise: no total, foram encontrados 12 dados envolvendo a concordância entre o sujeito e o predicativo do sujeito. Como ilustrado na tabela 43, a seguir, para facilitar a discussão, esses casos foram divididos em dados cujo núcleo do predicativo é um adjetivo e dados cujo núcleo é uma forma participial convertida em adjetivo.

CONCORDÂNCIA : SUJEITO-PREDICATIVO DO SUJEITO	
Adjetivos	5
Formas participiais convertidas em adjetivos	7
TOTAL	**12**

Tabela 43 – Concordância: Sujeito-Predicativo do Sujeito

Dado que essas categorias gramaticais não flexionam em número e gênero no texto de origem, em inglês, a tradução pode ser mais laboriosa. No caso dos predicativos com núcleo adjetival, foram encontrados dados com o sujeito explícito em PB, como (245b), em que o traço de gênero não é o adequado. Também foram encontrados dados com sujeito nulo em PB, como em (246b). A partir da observação dos traços do verbo "Estamos" é possível afirmar que o sujeito é plural, logo o predicativo foi incorretamente traduzido em (246b).

(245)

a. *if [the removal reason] is **unclear***
b. *se [a razão de remoção] não estiver **claro***
c. se [a razão de remoção] não estiver **clara***

(246)

a. *[We] are now **compatible** with...*
b. *[-] Estamos agora **compatível** com...
c. [-] Estamos agora **compatíveis** com...

O excerto (247b) exemplifica um erro de concordância de gênero em que o predicativo tem como núcleo uma forma participial convertida em adjetivo (uma passiva adjetival). O sujeito é explicíto, mas existe um erro de concordância dentro do SN que ocupa esta posição: o artigo definido não compartilha o traço de gênero do núcleo do SN. O mesmo erro de concordância existe entre o sujeito e o predicativo do sujeito. A proposta apresentada em (247c) seria a adequada neste caso.

(247)

a. *[the account] is **closed***
b. *[o conta] está **fechado***
c. [a conta] está **fechada***

Também foram encontrados 4 casos que envolvem construções passivas verbais com particípio passivo. Retomaremos aqui o dado já apresentado em (240), que repetimos em (248) devido à sua complexidade quando se considera também a tradução do particípio passivo, que, neste caso, mantém a sua categoria de verbo, dado que se trata de uma oração passiva. Como já foi afirmado acima, há a possibilidade de tradução do verbo "pode" em concordância semântica, no plural (248bi), ou em concordância gramatical, no singular (248bii), com o núcleo do SN, "maioria". Contudo, quando se observa o particípio

passivo dos excertos traduzidos pela empresa (248b), em (248bi) ele concorda em gênero e número e em (248bii) ele concorda somente em gênero com o nome "problemas", que é núcleo do SN inserido no SP. Na verdade, para que os traços dos particípios passivos estejam adequados em PB, seria necessária a correção da flexão do particípio de acordo com os traços de "maioria" (248ci) ou de acordo com os traços do SN mais encaixado (248cii), em que se estabelece uma concordância semântica quer no que diz respeito ao verbo "poder" quer no que diz respeito ao particípio.

(248)

a. *[Most visual issues] can be **resolved***
b.
   i. [A maioria dos problemas visuais] podem ser **resolvidos**
   ii. [A maioria dos problemas visuais] pode ser **resolvido**
c.
   i. [A maioria dos problemas visuais] pode ser **resolvida**
   ii. [A maioria dos problemas visuais] podem ser **resolvidos**

Para finalizar, a tabela 44 apresenta as contagens dos dados envolvendo a concordância Sujeito-Verbo, Sujeito-Predicativo do Sujeito e Sujeito-Particípio Passivo segundo os tipos de traços ou segundo os casos especiais envolvidos nas estruturas em PB. Esses casos especiais já foram discutidos sucintamente em exemplos anteriores e foram assinalados com as expressões "Flexão de Número" (a compra / s **será** deduzida, 241b), "Envolvendo Concordância Semântica" (A maioria dos problemas visuais **pode**(**m**) **ser**, 248b) e "Variação em PB" (ajudar viajantes a **programar**, 244b). Nesses dois últimos casos, não é possível falar efetivamente de erro de concordância, mas eles foram considerados na análise da presente pesquisa, pois refletem fenômenos interessantes do PB. Nesses casos, os parágrafos anteriores procuraram explicitar as dificuldades nas traduções encontradas nos dados e forneceram um espaço de discussão acerca desses fenômenos.

TIPOS DE CONCORDÂNCIA COM O SUJEITO			
Sujeito-Verbo	***Concordância de Número***	6	12
	Flexão de Número	1	
	Envolvendo Concordância Semântica	2	
	Variação em PB	3	
Sujeito-Predicativo do Sujeito	***Concordância de Gênero***	5	12
	Concordância de Número	5	
	Concordância de Número e Gênero	1	
	Flexão de Número	1	
Sujeito-Particípio Passivo	***Concordância de Gênero***	1	4
	Concordância de Número	1	
	Envolvendo Concordância Semântica	2	
TOTAL			**28**

Tabela 44 – Tipos de concordância com o sujeito

6.2.1.4 Problemas de concordância em outras estruturas

Foram encontrados 12 dados envolvendo concordância em outros tipos de estruturas em PB. Por serem casos pontuais e terem um baixo número de ocorrências, esses casos foram reunidos na presente subseção e contabilizados na tabela 45 a seguir. Quanto aos traços envolvidos nessas estruturas somente o dado com o particípio absoluto envolvia os dois traços de gênero e número em simultâneo. Já os dados com clíticos apresentam um maior número de erros com o traço de gênero e os dados com predicativos do objeto direto apresentam um maior número de erros com o traço de número.

CONCORDÂNCIA EM OUTRAS ESTRUTURAS		
Pronome Relativo	***Número***	1
Particípio Absoluto	***Gênero e Número***	1
Clítico	***Gênero***	3
	Número	2
Predicativo do Objeto Direto	***Gênero***	1
	Número	4
TOTAL		**12**

Tabela 45 – Concordância em outras estruturas

O exemplo seguinte foi o único envolvendo a concordância entre o pronome relativo e seu antecedente. A concordância em (249b) está incorreta, pois o núcleo do antecedente "modos" é plural, mas o pronome relativo "o qual" está no singular. Logo, a tradução apresentada em (249c) é a adequada.

(249)

a. *to see [the available payment modes for your country]* ***that*** *you can pay with*
b. *para ver [os modos de pagamento disponíveis para o seu país] com **o qual** pode pagar
c. para ver [os modos de pagamento disponíveis para o seu país] com **os quais** pode pagar

Os erros nos traços de gênero e/ou número dos clíticos encontrados nos dados só foram identificáveis através da relação entre o clítico e o seu antecedente, que permite construir a cadeia de referência correta. O clítico "lo" se refere a "créditos da (COMPANY M)", SN que ocorre como antecedente. Em PB, deve existir um compartilhamento de traços entre esses dois elementos para que a cadeia referencial seja construída adequadamente, como em (250c).

(250)

a. *If you have [(COMPANY M) credits] you may also use* ***it***
b. *Se você tem [créditos da (COMPANY M)], você também pode usá-**lo**
c. Se você tem [créditos da (COMPANY M)], você também pode usá-**los**

O excerto a seguir exemplifica um caso de concordância entre o objeto direto e o predicativo do objeto direto. Aqui também, o particípio em inglês não possui traços de gênero e número, mas em PB o particípio deve concordar com o núcleo do constituinte que modifica. No exemplo (251), o núcleo "restrições" é feminino e plural, logo só a tradução proposta em (251c) é adequada. Talvez uma das dificuldades para a tradução desse particípio seja a distância entre ele e o núcleo a que se refere.

(251)

a. *…that you have [any bank or international payment restrictions to (COMPANY M)]* ***removed****.*
b. *…que você tenha [restrições de pagamento bancárias ou internacionais à (COMPANY M)] **removida**.
c. …que você tenha [restrições de pagamento bancárias ou internacionais à (COMPANY M)] **removidas**.

Foi encontrado um único dado com o particípio absoluto, apresentado no exemplo (252) a seguir. A forma participial "arquivado" em (252b) se refere à expressão "as cobranças", por isso deveria apresentar os mesmos traços de gênero e número. Por isso, a sugestão (252c) seria a mais adequada. Tendo em vista o comportamento dos particípios absolutos ser muito peculiar, preferiu-se não aprofundar ainda mais neste tema na presente pesquisa.

(252)

a. *[Chargebacks] can take up to 45 days to get resolved once **filed***
b. *[as cobranças] podem demorar até 45 dias para serem resolvidas uma vez **arquivado***
c. [as cobranças] podem demorar até 45 dias para serem resolvidas uma vez **arquivadas**

6.2.1.5 Concordância e coesão referencial nos dados recolhidos

Como já foi possível verificar em alguns dos exemplos apresentados, o compartilhamento adequado dos traços de concordância é muito importante para a manutenção das relações de coesão textual (ver, por exemplo, o caso dos clíticos em (250)). Alguns dados já demonstraram a retoma dos traços de gênero, número e pessoa. Os próximos parágrafos discutem alguns aspectos relacionados com a concordância estabelecida pragmaticamente e a maneira como a ferramenta utilizada na empresa possibilita a resolução desses problemas.

O exemplo (253), já citado em (242), será retomado aqui para explicitar essas relações discursivas. Como já foi referido, os traços do verbo "ficar" podem ser retomados anaforicamente através do verbo "Adicione", pois o sujeito nulo de "ficar" retoma anaforicamente o sujeito nulo de "Adicione". Porém, para identificar os traços do particípio, é necessário ter informações sobre o contexto em que foi produzido o excerto. Considerando-se somente o texto desse dado é impossível descobrir o gênero do destinatário. Logo, as duas opções (253b) e (253c) são possíveis, mas uma delas pode estar incorreta quando inserida no contexto discursivo.

(253)

a. *[-] Add (...) to reduce the likelihood of [-] getting **locked out** of your account*
b. [-] Adicione (...) para reduzir a probabilidade de [-] ficar **bloqueado** na sua conta
c. [-] Adicione (...) para reduzir a probabilidade de [-] ficar **bloqueada** na sua conta

De modo geral, o anotador e o editor da Unbabel têm acesso a mais contexto, como os nomes do remetente e do destinatário nos espaços dedicados às assinaturas. Esses nomes são anonimizados pela empresa, mas mantêm os traços de gênero e número do texto original, permitindo assim a observação da concordância. Logo, um exemplo como (253) foi considerado como um caso de concordância estabelecida pragmaticamente na presente pesquisa, mas é, provavelmente, um caso de concordância estabelecidada anaforicamente para o editor e para o anotador, pois eles possuem mais contexto ao utilizarem a ferramenta. O excerto apresentado em (254b) é outro exemplo muito comum desse tipo de erro: o clítico "lo" se refere ao destinatário da mensagem. Nesse caso, foi possível observar a partilha de traços de gênero e número com o antecedente na presente pesquisa, pois o nome a qual ele se referia foi selecionado pelo anotador juntamente com o clítico. Esse nome, anonimizado manualmente com a expressão "(PERSON'S NAME F)", possui gênero feminino, logo a proposta (254c) é a mais adequada.

(254)

a. *I would be happy to help.*
b. *Terei muito prazer em ajudá-**lo**.
c. Terei muito prazer em ajudá-**la**.

6.2.1.6 Síntese dos problemas de concordância encontrados nos dados

A tabela 46 resume em números os dados mencionados na seção 6.2.1, tendo em conta os traços de concordância envolvidos nas estruturas. Não foi encontrado nenhum dado envolvendo erro no compartilhamento do traço de pessoa. Foram encontrados 62 erros envolvendo o traço de gênero, 39 envolvendo o traço de número e 13 envolvendo os traços de número e gênero em simultâneo. A categoria "Outros casos" se refere aos dados inseridos como erro de "Flexão de Número", "Envolvendo concordância semântica" e "Variação em PB"; nesses casos, não é possível falar propriamente de erro de concordância, por isso foram separados dos outros dados. Também foram inseridos nessa última linha da tabela os erros de "Concordância de Número ou Gênero" e de "Concordância de Número ou Não há erro" devido à impossibilidade em determinar com certeza quais traços são os mais adequados para os adjetivos "bancários" e "internacionais" no excerto "quaisquer restrições de pagamento bancários ou internacionais", apresentado em (237b).

PROBLEMAS ENVOLVENDO CONCORDÂNCIA	
Concordância de Pessoa	0
Concordância de Gênero	62
Concordância de Número	39
Concordância de Número e Gênero	13
Outros casos	18
TOTAL	**132**

Tabela 46 – Problemas envolvendo concordância

6.2.2 Ordem de palavras

Os problemas de ordem de palavras encontrados nos dados fornecidos pela empresa foram divididos em dois grupos: Internos ao SN e Internos à Frase. Foram encontrados 33 dados envolvendo a ordem entre possessivos, modificadores, complementos e o núcleo do SN. Quanto aos constituintes da frase, foram encontrados 42 dados envolvendo os complementos diretos e indiretos, clíticos e não-clíticos, e os advérbios modificadores de frase ou predicado. Como será visto mais adiante, muitos casos não constituem efetivamente erro de ordem, mas foram mantidos na presente análise, pois são interessantes para a discussão acerca das características do PB e das principais dificuldades na tradução e anotação envolvendo a ordem de palavras. Esses casos foram assinalados com as expressões "Possivelmente não natural" e "Variação em PB". Adicionalmente, será dada atenção especial aos clíticos e advérbios, tendo em vista as suas peculiaridades em PB.

DADOS ENVOLVENDO *ORDEM DE PALAVRAS*	
Internos ao SN	33
Internos à Frase	42
TOTAL	**75**

Tabela 47 – Dados envolvendo *Ordem de Palavras*

Antes de iniciar a discussão acerca dos problemas de ordem encontrados nos dados recolhidos, também nesta seção foi considerado importante fazer uma descrição suscinta acerca das principais características da ordem das palavras em inglês, buscando assim

observar as principais diferenças entre a LC e de partida, bem como compreender as possíveis razões por trás de certas escolhas feitas pelos editores e anotadores da empresa. Essa descrição geral do funcionamento da ordem em inglês será feita na subseção 6.2.2.1. As subseções 6.2.2.2 e 6.2.2.3 são dedicadas à análise dos dados envolvendo a ordem dos elementos internos ao SN e internos à frase, respectivamente. Finalmente, em 6.2.2.4 será apresentada uma síntese suscinta dos problemas de ordem apresentados nas citadas subseções.

6.2.2.1 Questões sobre ordem de palavras em inglês

Ao observar a estrutura de frases declarativas simples, Greenbaum (1996: 59), no seguimento de outros autores, afirma que, em sua estrutura básica, o inglês possui dois elementos obrigatórios: o sujeito e o verbo, ambos ocorrendo nesta ordem SV. A presença dos outros elementos da frase, como predicativos (P) e complementos (O), depende das propriedades de seleção do verbo. Esses elementos são inseridos após o verbo, constituindo assim a ordem SVO. Esse autor aponta cinco estruturas básicas em inglês, representadas abaixo com excertos retirados do texto do autor. Ao observar esta lista de exemplos da ordem básica em inglês (255), o caso que mais se distingue do PB é o (255c), pois em PB o objeto direto geralmente precede o indireto e em inglês é o complemento indireto que geralmente se posiciona logo após o verbo.

(255)

a. Sujeito – Verbo (SV): *My glasses (S) have disappeared (V).*
b. Sujeito – Verbo – Objeto Direto ou Indireto (SVO): *Our country (S) is absorbing (V) many refugees (O).*
c. Sujeito – Verbo – Objeto Indireto – Objeto Direto (SVOO): *I (S) am sending (V) you (O) an official letter of complaint (O).*
d. Sujeito – Verbo – Predicativo do Sujeito (SVP): *The water-bed (S) was (V) very comfortable (P).*
e. Sujeito – Verbo – Objeto Direto – Predicativo do Objeto (SVOP): *(S) have made (V) my position (O) clear (P).*

(Greenbaum, 1996: 71)

Essa ordem básica pode ser modificada em diversos contextos, como aponta Greenbaum (1996: 74). As orações interrogativas (256a) e as exclamativas (256b) apresentadas a seguir são exemplos de frases que podem se apresentar com outras estruturas, diferentes da ordem SVO. Preferiu-se não desenvolver essa questão na presente subseção, tendo em vista os dados fornecidos serem majoritariamente frases declarativas, seguindo geralmente as regras já citadas no parágrafo anterior.

(256)

a. Interrogativa: *Why are you sad?*
b. Exclamativa: *How silly I am!*

Tendo em vista os dados fornecidos, é interessante explicitar o funcionamento da ordem na voz passiva em inglês. Primeiramente, em inglês, existe uma construção chamada de duplo objeto, isto é em que os dois complementos são aparentemente complementos diretos. Para exemplificar, retomando o exemplo (255c) a palavra "you" exerce a função de complemento indireto, mas é aparentemente um complemento direto na frase (257a), considerando-se a ausência de preposição. Como aponta Greenbaum (1996: 65), o objeto indireto geralmente pode ser parafraseado por uma expressão introduzida por preposição, seguindo o objeto direto nesse caso, como exemplifica "*to you*" em (257b).

(257)

a. *I am sending **you** [an official letter of complaint].*
b. *I am sending [an official letter of complaint] **to you**.* (com modificações nossas)

(Greenbaum, 1996: 66)

Em inglês, qualquer um dos complementos podem ser elevados a sujeito na passiva. O objeto indireto da frase ativa pode se tornar o sujeito da ativa, como exemplifica "you" em (258a). Também o objeto direto da ativa pode ser elevado a sujeito da passiva, como o SN "*an official letter of complaint*" em (258b) e (258c). Nesse último caso, Greenbaum (1996: 66) aponta que geralmente o SP correspondente substitui o objeto indireto, como demonstra "*to you*" em (258c), mas a opção sem preposição também é possível (258b). Já o PB não dispõe da construção de duplo objeto, pelo que o único element que pode ocupar a posição de sujeito numa passive é o complemento direto típico da ativa (cf. seção 5).

(258)

a. ***You** are being sent [an official letter of complaint].*
b. *[An official letter of complaint] is being sent **you**.*
c. *[An official letter of complaint] is being sent **to you**.*

(Greenbaum, 1996: 66)

Segundo Greenbaum (1996: 70), em inglês, os advérbios e adjuntos adverbiais introduzem informações contextuais importantes na frase e podem ser constituintes

opcionais ou obrigatórios. Para exemplificar, a expressão "*last night*" é opcional em (259a), mas obrigatória em (259b).

(259)

a. *I had a really good supper **last night**.*
b. *Our committee meeting was **last night**.*

(Greenbaum, 1996: 70)

Relativamente à ordem, os advérbios podem ocupar diversas posições dentro da frase. As frases apresentadas abaixo, retiradas do texto do autor, exemplificam esses elementos posicionados no início da frase (260a), no final da frase (260b) e entre o sujeito e o verbo (260c). No caso de alguns advérbios, o verbo segue-os se for um verbo principal (260c) e precede-os se for um verbo auxiliar ou o copulativo *to be* (260d). Encontram-se nesta situação os advérbios de frequência e os de intensidade, como exemplificam os excertos abaixo:

(260)

a. ***In the summer** you can take a car and four people for a hundred and twenty pounds.*
b. *You need a lot of strength **in the right hand**.*
c. *He **merely** shrugged his shoulders.*

(Greenbaum, 1996: 70 & 59)

d. *He was **merely** a young boy.*

Quanto à ordem dentro do SN, os determinantes se posicionam antes do núcleo em inglês, como "*A*" em (261a). Também é possível a coocorrência de diferentes especificadores, como o possessivo "*our*" e o quantificador "*all*" em (261b). O SN também pode conter mais de um modificador: em (261a) o núcleo "*citizen*" é modificado por "*second-class*", posicionado antes do núcleo, e "*of his own clan*", posicionado após o núcleo.

(261)

a. ***A** second-class citizen of his own clan.*
b. *In the initial sorties **all our** aircraft have returned safely.*

(Greenbaum, 1996: 209, com grifos nossos)

As diferentes relações de dependência entre os elementos do SN são evidenciadas por Greenbaum (1996: 209-210). Segundo esse autor, o núcleo pode ser acompanhado por

mais de um modificador pré-nominal, mas as relações sintáticas entre esses elementos e o núcleo podem ser diferentes. Para exemplificar, em (262a), "*spotted*" modifica "*hyenas*" e "*female*" modifica toda a unidade "*spotted hyenas*". Já em (262b), ambas as palavras "*large*" e "*aggressive*" modificam "*females*".

(262)

a. *[female [spotted **hyenas**]]*
b. [[*large*] [*aggressive*] ***females***]

(Greenbaum, 1996: 209-210, com grifos nossos)

Essa mesma diferença de relações pode ocorrer com modificadores pós-nominais, como também evidencia esse autor (Greenbaum, 1996: 210). Em (263a), o núcleo é "*corporation*" e os dois modificadores pós-nominais são "*I have worked for*" e "*where this has been a problem*". Nesse caso, o segundo modificador pós-nominal está modificando toda a expressão "*corporation I have worked for*". O caso de (263b) é diferente, pois os dois modificadores "*dated 22nd March 1990*" e "*for £43.13*" modificam ambos o núcleo "*invoice*".

(263)

a. *I think it is a pity that LB is the only major [**corporation** I have worked for [where this has been a problem]].*
b. *We could not trace the [**invoice** [dated 22nd March 1990] [for £43.13]].*

(Greenbaum, 1996: 210)

Quanto à posição mais típica dos modificadores em função da sua categoria gramatical, Greenbaum (1996: 217-218) afirma que adjetivos são tipicamente inseridos antes do núcleo, como exemplifica (264a). Mas também particípios (264b), gerúndios (264c), SNs (264d) são possíveis nesta posição, como exemplificam os excertos assinalados a seguir:

(264)

a. Adjetivo: *I hope that you shouldn't start somebody on **life-long anti-hypertensive** therapy based upon one **single** blood-pressure measurement.*
b. Particípio: *The results of that in pollution and **wasted** natural resources...*
c. Gerúndio: *But I hope to throw the net further in the **coming** weeks and...*
d. Nome: *One of the **consortium** members...*

(Greenbaum, 1996: 217-218)

Como já foi mencionado na seção 6.3.2.1, a posição após o núcleo do SN é tipicamente ocupada por complementos ou modificadores. Segundo Payne (2010: 5), os modificadores pós-nominais tendem a ser mais pesados e incluem geralmente SP ou orações relativas. Os excertos assinalados a seguir exemplificam um SP modificador (265a) e um SP complemento (266a), bem como orações com função de modificador (265b) e complemento (266b) do núcleo do SN.

(265)

a. *an iota* ***of truth***
b. *The last time* ***I saw her face***

(Payne, 2010: 5)

(266)

a. *He treats patients* ***with head injuries****.* (Payne, 2010: 9)
b. *the idea* ***that Mary is coming****.*

Como aponta Payne (2010: 241-242), o uso de modificadores adjetivais após o núcleo do SN é mais restrito, mas também é possível nos seguintes contextos: a) adjetivos como relativas truncadas, ou seja, adjetivos que restringem o núcleo da mesma maneira que uma oração relativa; b) adjetivos acompanhando pronomes indefinidos; e c) adjetivos com um complemento próprio. No caso de (267a), por exemplo o sentido do adjetivo "*present*" é o mesmo da oração relativa "*who were present*", por isso sua identificação como relativa truncada. Em (267b), o adjetivo "wicked" acompanha o pronome indefinido "*something*" e em (267c) o adjetivo "*desirous*" licencia o complemento "*of a truly elegante abode*".

(267)

a. *Those members* ***present*** *refused to select a candidate.*
b. *Something* ***wicked*** *this way comes.*
c. *...or a family [****desirous*** *[of a truly elegant abode]]...*

(Payne, 2010: 241-242)

A partir do exposto, é possível concluir que, em geral, a estrutura frásica e a organização dos elementos dentro do SV em PB e em inglês não são muito distintas. Os maiores obstáculos para a tradução residem nas diferenças no interior do SN. Em primeiro lugar, certas categorias gramaticais são modificadoras típicas do inglês, mas não são podem ser modificadoras do SN em PB, sendo necessário fazer mudar essas categorias; é o caso dos SNs que modificam expressões nominais em inglês, mas não em PB. Em

segundo lugar, enquanto em PB a posição típica dos modificadores adjetivais é após o núcleo do SN, havendo adjetivos que podem ocupar ambas as posições, muitas vezes com alteração de significado, e sendo escassos os adjetivos exclusivamente pré-nominais (cf. seção 5), em inglês a posição típica é antes desse núcleo, também dificultando a tradução da ordem desse tipo de palavra. Em terceiro lugar, dado que modificadores como os adjetivos e os particípios não apresentam marcas de flexão de gênero nem de número, pode ser difícil identificar corretamente as relações de entre os elementos dentro do SN e traduzir em PB numa ordem que reflita adequadamente essas relações.

Os próximos parágrafos seguem uma organização semelhante à subseção 6.2.1 anterior, ou seja, foram inseridas algumas tabelas com as contagens segundo as categorias gramaticais e os exemplos apresentados nesta subseção foram retirados dos dados fornecidos pela empresa e se organizam da seguinte maneira: o item (a) contém o texto na LP; o item (b) apresenta a tradução na LC feita através do processo de tradução da empresa; os itens (c) e (d) são propostas de tradução ou exemplos de tradução que podem auxiliar a análise que será feita na subseção.

6.2.2.2 Ordem de palavras internamente ao SN

Nos dados coligidos, foram encontrados 33 casos de erros envolvendo a ordem de palavras no interior do SN. Na maior parte desses dados, o elemento na posição incorreta é um modificador do SN, correspondendo a 55% dos casos. Em seguida, na segunda posição estão os dados com complementos do SN na posição incorreta: 12 argumentos do nome foram mal posicionados na tradução feita pela Unbabel. No caso dos especificadores, somente foram encontrados 2 erros com o possessivo.

ORDEM DE PALAVRAS: INTERIOR DO SN	
Modificadores do SN	19
Complementos do SN	12
Possessivos	2
TOTAL	**33**

Tabela 48 – Ordem de palavras: Interior do SN

Nos dois casos com possessivo o erro de ordem ocorre juntamente com erro de concordância: há erro de ordem e concordância de número em (268) e erro de ordem e

concordância de gênero em (269). Em (268b), a palavra "DJ" foi interpretada como núcleo do SN. Porém, como é possível observar em (268a) o núcleo do SN é "*products*". Considerando-se o contexto discursivo da frase, disponibilizado nas instruções enviadas pelo cliente da Unbabel, é possível saber que "*DJ*" é o nome próprio que designa o produto, pelo que o possessivo não especifica apenas esse nome, mas antes o grupo *DJ products*. Também em (269) há uma interpretação inadequada do núcleo do SN: no original em inglês o possessivo "*your*" especifica toda a expressão "*(COMPANY M) Account*" (269a), mas na tradução feita pela empresa o possessivo "seu" especifica somente a expressão anonimizada "(COMPANY M)".

(268)

a. ***our*** *DJ products*
b. *produtos do **nosso** DJ
c. **nossos** produtos DJ

(269)

a. ***your*** *(COMPANY M) Account*
b. *as Conta do **seu** (COMPANY M)
c. a **sua** Conta (COMPANY M)

Quanto aos modificadores do SN, é interessante fazer um paralelo entre as categorias gramaticais no texto de partida e a tradução correspondente na LC, pois há diferenças entre as estruturas do SN em inglês e em PB. Como será visto mais adiante, além da ordem, há casos em que também a categoria gramatical do elemento no texto de chegada está incorreta, visto que nomes comuns e formas gerundivas podem modificar diretamente o núcleo do SN em inglês, mas não em PB. Tendo em vista essas diferenças, é compreensível que os dados em que se verifica mais erros internamente ao SN são aqueles em que no texto de partida ocorre um nome como modificador de outro nome: foram encontrados 7 dados em que o nome estava na posição incorreta. O adjetivo é a segunda categoria com mais erros: há 6 erros de ordem envolvendo esse tipo de palavra. Isso também é previsível, considerando que a posição típica dos adjetivos dentro do SN é muito diferente quando se compara as duas línguas. Dentro dos elementos que funcionam como modificadores do SN, também foram encontrados 5 erros envolvendo a forma participial convertida em adjetivo e 1 envolvendo a forma gerundiva convertida em adjetivo.

ORDEM DE PALAVRAS: MODIFICADORES DO SN		
No texto de partida	**No texto de chegada**	
Adjetivo	***Adjetivo***	6
Forma participial	***Forma participial***	5
Forma gerundiva	***Nome***	1
Nome	***Nome***	7
TOTAL		19

Tabela 49 – Ordem de palavras: Modificadores do SN

Nos erros com adjetivos, foram encontrados dados em que o adjetivo era o único modificador, como em (270a). Na expressão apresentada em (270b) a ordem foi replicada de maneira literal: o adjetivo foi mantido na posição pré-nominal (e não foi feita a concordância entre os traços do modificador e do núcleo). A expressão apresentada em (270c) seria a adequada em PB, pois o adjetivo em causa é obrigatoriamente pós-nominal. Esse tipo de erro foi assinalado como "Ordem de Palavras e Concordância de Número" na presente pesquisa.

(270)

a. ***secret*** *boards*
b. ***secreto** painéis
c. painéis **secretos**

O exemplo apresentado em (271) já foi citado na seção 6.3.2, mas será repetido aqui tendo em vista a complexidade das relações presentes. No sintagma do texto original (271a), há uma alternância entre os adjetivos "*minimum/maximum*", ambos modificando o núcleo "*durations*". A tradução apresentada em (271b) não é agramatical, mas não é fiel ao sentido do texto de partida, pois há uma alternância entre dois grupos nominais com núcleos distintos ("treino" e "duração"). As propostas de tradução apresentadas em (271c) e (271d) seriam mais adequadas, tendo em vista o sentido do texto de partida. Durante a presente pesquisa, a palavra "mínimo" foi assinalado como "Ordem de Palavras e Concordância de Gênero e Número" e a palavra "máximo" foi assinalada como "Ordem de Palavras e Concordância de Número".

(271)

a. *workout **minimum/maximum** durations*
b. ? treino **mínimo /** duração **máxima**
c. durações **mínimas/máximas** de treino
d. durações **mínimas** e **máximas** de treino

No exemplo (272a) a seguir, há mais uma vez a presença do símbolo "/" possibilitando uma alternância. Em PB, o particípio adjetival não pode ocorrer antes do nome e, caso também exista um modificador adjetival, a posição mais adequada é após esse elemento, como em (272c) e (272d). A tradução do particípio encontrada nos dados (272b) contém erro de ordem juntamente com erro de concordância de gênero e número.

(272)

a. *not e-money/**pre-paid** financial institutions*
b. *não e-money / **pré-pago** instituições financeiras
c. não e-money / instituições financeiras **pré-pagas**
d. não e-money ou instituições financeiras **pré-pagas**

Também em (273a) o particípio modifica o núcleo juntamente com outros elementos: o nome "*(COMPANY F)"* e o adjetivo "*return*". Tendo em vista o nome não poder modificar sozinho o núcleo do SN, o nome "(COMPANY F)" foi inserido em um SP (273b). O excerto traduzido pela empresa não é agramatical, mas a posição mais natural para o particípio em PB é antes do SP e após o adjetivo, como em (273c), tendo em vista esse particípio causar certa ambiguidade quando posicionado logo após "da (COMPANY F)". Outro caso semelhante é apresentado em (274b) em que a posição da forma participial "reconhecido" não é agramatical, mas pode causar certa ambiguidade: além da sua leitura como modificador de todo o grupo nominal "um banco do (COUNTRY M)", como está no original, há a leitura dessa palavra como modificador inserido no SP, modificando somente o país "(COUNTRY M)". A posição mais natural seria entre o núcleo e o SP modificador. Ambos os casos foram assinalados como "Possivelmente não natural" na análise da presente pesquisa, pois não há erro de ordem, mas são interessantes para a discussão acerca da posição dos elementos dentro do SN.

(273)

a. *your **prepaid** (COMPANY F) return label*
b. seu rótulo de retorno da (COMPANY F) **pré-pago**
c. o seu rótulo de retorno **pré-pago** da (COMPANY F)

(274)

a. ***recognized*** *(COUNTRY M) bank*
b. um banco do (COUNTRY M) **reconhecido**
c. um banco **reconhecido** do (COUNTRY M)

Como já foi visto, em inglês a forma "*-ing*", correspondente ao gerúndio em PB, pode modificar um grupo nominal (275a). Tendo em vista a impossibilidade desse elemento como modificador em PB, o constituinte que, em inglês, integra a forma gerundiva deve ser realizado como outra categoria sintática em PB. Ainda que esta alteração de categoria seja feita em muitos casos, há problemas na ordem de constituintes; por exemplo, na tradução (275b), o constituinte em questão se mantém em posição pré-nominal. Assim, o excerto apresentado em (275c) é o adequado, pois "*billing*" foi realizado como SP, "de cobrança", e inserido após o núcleo, pois essa é a posição apropriada para esse sintagma, quando inserido no SN.

(275)

a. *a billing ticket*
b. *um **cobrança** ingresso
c. um ingresso **de cobrança**

Como se pode inferir a partir da porcentagem de erros encontrados nos dados, a tradução envolvendo estruturas em que os nomes (ou SNs) modificam outros nomes são mais complexas e difíceis de traduzir em PB, não somente devido ao fato de os nomes não modificarem o núcleo do SN em PB, mas também porque o nome modificador, geralmente posicionado antes do núcleo em inglês, pode ser incorretamente interpretado como núcleo do SN e não como modificador. Para exemplificar, em (276b) e (277b), os nomes "empresa" e "cartão" estão na posição típica do núcleo, ou seja, logo após o especificador. No entanto, esses nomes são na verdade modificadores e deveriam ocorrer após o núcleo. Além disso, para que o sintagma seja adequado em PB, os SNs "*business*" e "*card*" devem ser realizados como SA, em (276c), ou como SP, em (277c). Tendo em vista a necessidade de mudança da categoria gramatical do elemento "empresa" em (276b), esse segmento foi assinalado com a expressão "Categoria gramatical".

(276)

a. *a **business** account*
b. *uma **empresa** conta
c. uma conta **empresarial**

(277)

a. *any **card** payments*
b. *qualquer **cartão** pagamentos
c. quaisquer pagamentos **com cartão**

Também foram encontrados nomes próprios mal posicionados dentro do SN nas traduções feitas pela empresa. Para exemplificar, em (278a), a expressão "*Settings*" é um nome próprio, referindo-se ao nome do ícone. Nesse caso, o nome comum "ícone" funciona como classificador e deve ser inserido antes do nome próprio, sendo a ligação entre esses dois elementos geralmente feita diretamente (cf. seção 5), como na proposta (278c).

(278)

a. *on the **Settings** icon*
b. no **Configurações** ícone
c. no ícone **Configurações**

No caso dos complementos do SN, também é interessante fazer um paralelo entre a categoria gramatical dos elementos no texto de partida e chegada, tendo em vista as já mencionadas diferenças entre a organização do SN em PB e em inglês (cf. seções 5 e 6.3.3.1). Nos dados fornecidos, foram encontrados 8 casos em que o nome complemento no texto de partida foi traduzido por outro nome no texto de chegada, faltando nesses casos, além da correção da ordem, a inserção do nome dentro de um SP ou a alteração da categoria gramatical desse elemento em PB, pois o nome não pode complementar sozinho o núcleo do SN. Por último, foram encontrados 4 casos em que um grupo nominal inteiro funcionava como complemento do núcleo do SN no texto de partida e foi traduzido por um grupo nominal correspondente em PB. Entretanto, como nos casos do nome complemento, também os grupos nominais devem ser inseridos em um SP quando exercem a função de complementos do SN.

ORDEM DE PALAVRAS: COMPLEMENTOS DO SN		
No texto de partida	**No texto de chegada**	
Nome	***Nome***	8
Grupo nominal	***Grupo nominal***	4
TOTAL		12

Tabela 50 – Ordem de palavras: Complementos do SN

A posição típica do núcleo do SN em inglês é no final desse sintagma, sendo mais comum encontrar os complementos à esquerda do núcleo. Já em PB, a posição mais típica do complemento é à direita do núcleo do SN, logo em (279b) as posições das palavras "assinatura" e "plano" devem ser invertidas. Além de estar mal posicionado dentro do SN, o complemento "assinatura" também deveria estar inserido dentro de um SP, por isso a proposta em (279c) seria a mais adequada. Em vista disso, esse tipo de dado foi assinalado na presente análise com a expressão "Ordem de Palavras e Falta preposição".

(279)

a. *a new **subscription** plan*
b. um novo **assinatura** plano
c. um novo plano **de assinatura**

Já no dado apresentado em (280b) foi assinalado nesta pesquisa como contendo, além do erro de ordem de palavras, erro de categoria gramatical e concordância de número ou erro de falta de preposição. Isso porque há duas possibilidades de tradução para "*bank*" e a sua tradução depende então das escolhas do tradutor: pode ser um SA em (280c); ou um SP em (280d). Caso seja escolhida a tradução por um SA, é necessário também inserir a devida concordância entre os traços do núcleo desse SA e do núcleo do SN. Por isso, esse dado foi assinalado como contendo erro de "Ordem de Palavras, Categoria gramatical e Concordância de número **ou** Ordem de Palavras e Falta Preposição".

(280)

a. *to **bank** inquiries*
b. *ao **banco** inquéritos
c. aos inquéritos **bancários**
d. aos inquéritos **do banco**

Como já foi mencionado anteriormente, a posição pós-nominal é a mais típica para os complementos. Considerando-se essa afirmação, a ordem dos elementos na tradução apresentada em (281b) é muito próxima à estrutura do texto original e a tradução proposta em (281c) seria a mais natural. Porém, quando se considera o PB, não é raro encontrar a expressão "Vídeo chamada", como apresentado em (281b), em textos que tratam de ferramentas tecnológicas. Por isso, nesse caso é necessário mais contexto discursivo para determinar se a expressão (281b) é inadequada. Em vista disso, esse dado foi considerado como um caso interessante para a discussão da ordem no SN e foi assinalado como "Possivelmente não natural".

(281)

a. *'**Video** call'*
b. '**Vídeo** chamada'
c. 'Chamada **de vídeo**'

O excerto a seguir exemplifica a complementação feita por um grupo nominal: em (282a) o bloco [*Full page*] complementa o núcleo [*screenshot*]. Primeiramente, a palavra "*screenshot*" não pode ser traduzida por um só termo, logo ela foi traduzida pela expressão [imagem de captura de tela] em (282b). Em segundo lugar, a posição dos elementos no excerto (282b) está incorreta, pois o bloco complementador [página completa] deve ser posicionado após a expressão que ele complementa, como em (282c). Em terceiro lugar, esse grupo nominal com função de complemento, como no caso dos nomes complementos, deve ser inserido num SP, por isso também falta preposição em (282b). Há outros erros em (282b), como a concordância no determinante "*o*" e a omissão do pronome relativo "*que*": esses erros foram corrigidos na proposta de tradução, mas não são relevantes para a discussão do presente parágrafo. Apesar do bom posicionamento do grupo nominal na proposta de tradução, a expressão em (282c) continua causando estranhamento devido à quantidade de SPs existentes dentro do mesmo SN. Esse tipo de erro foi assinalado nesta análise como "Ordem de Palavras e Falta de Preposição".

(282)

a. *[**Full page**] [screenshot] of the error message you receive*
b. *[**Página completa**] [imagem de captura de tela] do mensagem de erro você recebe
c. ? [Imagem de captura de tela] [**da página completa**] da mensagem de erro que você recebe

Para finalizar esta subseção, a tabela 51 a seguir lista os tipos de problemas de ordem internamente ao SN encontrados nos dados fornecidos pela Unbabel, incluindo os três grupos citados na tabela 48: possessivos, modificadores e complementos. Foram contabilizados 11 casos em que o problema envolve somente a ordem de palavras. A maior parte dos dados possui problemas na ordem de palavras juntamente com outro tipo de erro, totalizando 18 casos: 6 dados possuem erros de ordem e concordância (271b, treino **mínimo /** duração **máxima**); 9 envolvem a falta de preposição após a correção da ordem (279b, um novo **assinatura** plano); e 1 dos dados implica na mudança de categoria gramatical após a correção da ordem (276b, uma **empresa** conta). Vale lembrar que 2

dados foram assinalados com a expressão "Categoria gramatical e Concordância de Número ou Falta de Preposição", tendo em vista as suas peculiaridades (280c, aos inquéritos **bancários**; ou 280d, aos inquéritos **do banco**). Finalmente, 4 casos foram assinalados como "Possivelmente não natural" (274b, um banco do (COUNTRY M) **reconhecido**), dado que a sua classificação como erro depende do contexto discursivo, não disponível durante a presente pesquisa.

TIPOS DE PROBLEMAS DE ORDEM DE PALAVRAS NO SN			
Ordem de Palavras		11	
Ordem de Palavras + ...	***Concordância de Número***	3	18
	Concordância de Gênero	2	
	Concordância de Número e Gênero	1	
	Categoria gramatical	1	
	Categoria gramatical e Concordância de Número ou Falta de Preposição	2	
	Falta de preposição	9	
Possivelmente não natural		4	
TOTAL		**33**	

Tabela 51 – Tipos de problemas de ordem de palavras no SN

6.2.2.3 Ordem dos elementos na frase

Relativamente à ordem dos elementos dentro da frase, foram encontrados 40 dados interessantes para a análise, listados na tabela 52 a seguir. Os complementos diretos foram envolvidos em 4 desses dados quando não são clíticos e em 8 casos quando são clíticos. Quanto aos complementos indiretos, foi encontrado 1 dado envolvendo um elemento não clítico e encontrados 6 com clíticos. Também há 2 casos envolvendo a ordem de complementos oblíquos. Foram encontrados 21 problemas envolvendo a ordem de advérbios ou constituintes com valor adverbial.

ORDEM DE PALAVRAS: INTERIOR DA FRASE		
Complemento direto	***não clítico***	4
	clítico	8
Complemento indireto	***não clítico***	1
	clítico	6
Complemento oblíquo		2
Advérbios e constituintes com valor adverbial		21
TOTAL		**42**

Tabela 52 – Ordem de palavras: Interior da Frase

Nesta subseção, serão comentados problemas relativos à ordem de palavras na frase. Dado que a posição dos clíticos e dos advérbios assume particular importância em PB, quer pela frequência do erro quer pela ausência, no caso dos advérbios, de uma posição fixa, os dados que envolvem estes elementos foram separados em subseções específicas para eles: em 6.2.2.3.1, apresenta-se os problemas envolvendo complementos não clíticos; em 6.2.2.3.2 são analisados com mais detalhe os problemas relacionados com a posição dos complementos clíticos; e em 6.2.2.3.3 a posição dos advérbios encontrados nos dados fornecidos pela empresa.

6.2.2.3.1 Problemas envolvendo complemento não clíticos

Além do erro ligado à forma do verbo "ajuda" em (283b), há erro na posição do complemento direto "viajantes". Esse elemento foi inadequadamante inserido entre a preposição e o verbo infinitivo "programar". Por isso a sugestão apresentada em (283c) seria a adequada neste caso.

(283)

a. *to help **travellers** plan their trips*
b. *para ajuda a **viajantes** programar as suas viagens
c. para ajudar **viajantes** a programar as suas viagens

Como já foi tratado na seção 5, a ordem típica dos elementos frásicos em PB é SVO (Sujeito-Verbo-Objeto), sendo também possível posicionar esses elementos em outras partes da frase em PB. Considerando essas informações, o dado assinalado em (284) exemplifica um caso classificado como "Possivelmente não natural", pois não há agramaticalidade na tradução feita em (284b): apesar o complemento direto "o seguinte"

poder ocorrer após o complemento oblíquo "no questionário", a frase se torna mais natural se aquele seguir imediatamente o verbo, como no exemplo (284c).

(284)

a. *you indicated in the questionnaire* ***the following***
b. você indicou no questionário, **o seguinte**
c. você indicou **o seguinte** no questionário

O exemplo (285) é um caso interessante, pois a oração foi traduzida de maneira muito próxima à estrutura do texto original, causando estranhamento em (285b). Uma maneira de corrigir esse estranhamento, apresentada nas propostas de tradução (285c) e (285d), é transformar o complemento oblíquo "*for a company name*" em complemento direto em PB ("o nome da empresa") e transformar o complemento direto do original "*your credit card statement*" em complemento oblíquo em PB, através da introdução da preposição "em". Após essas mudanças, ambos os excertos com posições diferentes (285c) e (285d) são possíveis em PB, logo não há agramaticalidade na ordem da tradução (285b), feita na empresa, pois o complemento direto e o oblíquo podem trocar de posição na frase. Entretanto, semelhante ao caso (284c) apresentado anteriormente, a ordem apresentada em (285c) soa mais natural em PB, pois o complemento direto está posicionado logo após o verbo. Tendo em vista os problemas de ordem encontrados nesse dado levantarem questões interessantes para a análise feita na presente pesquisa, optou-se por assinalar esse caso com a expressão "Possivelmente não natural" e mantê-lo nas contagens.

(285)

a. *you can usually check [your credit card statement] [****for a company name****].*
b. ? [-] pode verificar [a declaração do seu cartão de crédito] [**para o nome da empresa**].
c. [-] pode verificar [**o nome da empresa**] [na declaração do seu cartão de crédito].
d. [-] pode verificar [na declaração do seu cartão de crédito] [**o nome da empresa**].

A ordem na voz passiva pode parecer emaranhada, pois não há alinhamento entre a função sintática e semântica. Além disso, como já foi observado na seção 6.2.2.1 acerca da ordem das palavras em inglês, nessa língua é possível construções com duplo objeto, em que dois elementos são aparentemente complementos diretos. Consequentemente, ambos os complementos "*you*" e "*$ (PRICE)*" em (286a) podem ocupar a posição de

sujeitos da passiva em inglês (cf. "*$ (PRICE) was immediately charged from you"* e "*you were immediately charged $ (PRICE)")*. Já em PB o único elemento que pode ocupar a posição de sujeito numa passiva é o complemento direto típico da ativa (cf. seção 5). Por isso, ao traduzir o excerto com uma ordem muito próxima ao original em (286b), a frase é considerada agramatical ou causa muito estranhamento para os falantes de PB, pois o elemento "você" é tipicamente o complemento indireto da frase ativa, mas ocupa a posição de sujeito da frase passiva em (286b). Tendo em conta essas informações, a ordem proposta em (286c) seria mais adequada, sendo necessário fazer a devida concordância entre os elementos "foram cobrados" e o traço singular ou plural do numeral anonimizado em "$ (PRICE)".

(286)

a. *[**you**] were immediately charged [$(PRICE)]*
b. */?? [**você**] foi cobrado imediatamente [$ (PRICE)]
c. foram cobrados imediatamente [**de você**] [$(PRICE)]

6.2.2.3.2 Problemas envolvendo complemento clíticos

Como já foi visto na discussão acerca dos clíticos apresentada na seção 5.2.1.2, existem regras acerca das possíveis posições para esses elementos e certas posições não são possíveis para os clíticos em PB. Considerando as descrições sobre a posição dos clíticos em PB, todos os dados envolvendo esses elementos foram assinalados como casos de "Variação em PB". Isso porque apesar de existirem posições mais adequadas para os clíticos em alguns dos dados fornecidos pela empresa, como será explicitado nos exemplos a seguir, não foram encontrados clíticos em posições decisivamente agramaticais. Como será visto nos exemplos explicitados nos parágrafos a seguir, há uma coexistência entre a tendência proclítica do PB e a o uso emaranhado da ênclise. Tendo isso em consideração, pode ser difícil determinar com certeza qual posição é a mais adequada para os clíticos encontrados. Por isso, algumas das ideias já apresentadas em 5.2.1.2 serão repetidas aqui para auxiliar na análise das possíveis posições dos clíticos nos exemplos selecionados.

A divisão apresentada na tabela 53 segue a proposta de Kanthack (2002), considerando-se que esses dois grupos de clíticos se comportam de maneiras distintas em PB: foram encontrados 12 dados envolvendo os clíticos do primeiro grupo — "me", "te", "se", "lhe" e variantes; e 2 dados envolvendo os clíticos do segundo grupo — "o" e variantes.

ORDEM DE PALAVRAS: CLÍTICOS	
Grupo 1 (“me”, “te”, “se”, “lhe” e variantes)	12
Grupo 2 (“o” e variantes)	2
TOTAL	**14**

Tabela 53 – Ordem de palavras: Clíticos

No caso dos clíticos do grupo 1, Kanthack (2002) afirma que esses clíticos podem surgir na posição inicial da frase, mesmo na ausência de elementos fonéticos à sua esquerda, como em (287b) e (288c). Também segundo Luís e Kaiser (2016), em PE a ênclise é exigida em frases simples de sujeito nulo, como em (287c) e (288b), mas a próclise é permitida em PB, como nas traduções em (287b) e (288c). Contudo, esses autores também ressaltam que a ênclise do clítico “se” é bastante favorecida em PB em posição inicial de frase, logo a opção (287c) parece ser a mais natural. A partir das observações desses autores, é possível perceber que em PB ambas as posições apresentadas são possíveis, sendo difícil determinar a melhor maneira de anotar esse erro durante o processo de anotação ou o modo mais adequada de orientar os anotadores e editores.

(287)

a. *Please also kindly ensure*
b. **Se** certifique também
c. Certifique-**se** também

(288)

a. *If you’d like to change to the monthly subscription, please let me know and I’ll be glad...*
b. Se você quiser mudar para a subscrição mensal, informe-**me** e terei todo o gosto...
c. Se você quiser mudar para a subscrição mensal, **me** informe e terei todo o gosto...

Em perífrases verbais do PB, segundo Luís e Kaiser (2015) e Galves (2001), o clítico do grupo 1 se liga em próclise ao verbo principal. Por isso, a proposta (289c) parece ser a mais natural em PB. Apesar de menos natural, também é possível encontrar em PB clíticos posicionados como no excerto (289b). Também esse tipo de dado levanta questões acerca de como anotar casos como o (289b) durante o processo de anotação feito na empresa, por isso dados como esse foram assinalados com a expressão “Variação em PB”.

(289)

a. *so we can give you a temporary passcode*
b. ? para que possamos dar-**lhe** uma senha temporária
c. para que possamos **lhe** dar uma senha temporária

No caso de sentenças preposicionadas com verbos não-finitos, o clítico de grupo 1 pode se posicionar antes (290c) ou depois (290c) do verbo, segundo Luís e Kaiser (2015). Já Galves (2001) afirma que no caso de frases com o infinitivo, o pronome se insere em próclise em sentenças iniciadas por preposição, logo a opção (290c) seria a mais natural em PB. Considerando as afirmações desses autores, a posição apresentada em (290b) não é agramatical em PB, logo, como nos exemplos anteriores, esse dado também foi assinalado com a expressão "Variação em PB".

(290)

a. *I'm really sorry for the situation and would like to keep you posted*
b. Sinto muito pela situação e gostaria de manter-**te** informada
c. Sinto muito pela situação e gostaria de **te** manter informada

Os clíticos de grupo 2 possuem um comportamento diferenciado: apesar da tendência proclítica, há contextos em que a ênclise é favorecida em PB. As sentenças de infinitivo não preposicionado são apontadas por Galves (2001) e Luís e Kaiser (2005) como contextos de favorecimento da ênclise. De acordo com Kanthack (2002: 124), os clíticos do segundo grupo não podem estar antes do verbo não-finito, por isso a agramaticalidade do exemplo (291d), eles também podem ocorrer antes do verbo finito, como em (291b), mas a posição mais adequada segundo essa autora é em ênclise ao verbo principal não-finito, como em (291c). Também nesse caso, por não existir agramaticalidade, mas sim uma naturalidade menor, na ordem proposta pela tradução feita na empresa, o dado (291b) foi assinalado com a expressão "Variação em PB".

(291)

a. *To help me further assist you*
b. ?? Para que eu **a** possa ajudar
c. Para que eu possa ajudá-**la**
d. *Para que eu possa **a** ajudar

No exemplo (292), a sentença apresenta, além do clítico de grupo 2, um verbo não-finito e a preposição "para". Segundo as ideias de Kanthack (2002: 117), nesse caso os clíticos de segundo grupo são licenciados na posição pós-verbal, em ênclise ao verbo não-

finito, como em (292c). Logo a ordem nessa proposta de tradução seria a mais adequada em PB. Já a tradução feita pela empresa (292b) causa muito estranhamento para os falantes do PB, mas ainda assim é possível encontrar sentenças em PB escrito com esse tipo de estrutura (cf. seção 5). Também nesse caso é difícil determinar de modo generalizado a existência de agramaticalidade na ordem em (292b), dificultando a uniformidade na anotação feita na empresa.

(292)

a. *to help you choose*
b. ?? para **o** ajudar a escolher
c. para ajudá-**lo** a escolher

6.2.2.3.3 Posição dos advérbios nos dados recolhidos

Na presente subseção, serão observados alguns dos 20 dados envolvendo a ordem dos advérbios e dos 2 dados envolvendo outros constituintes com valor adverbial. Como será visto mais adiante, com exceção de um dado em que há uma mudança drástica do sentido do texto original, esses dados não foram classificados como erro, mas foram assinalados com a expressão "Possivelmente não natural" na presente pesquisa. Isso ocorreu devido às características das estruturas envolvendo advérbios em PB: esses elementos podem se posicionar em diversas partes da frase, como já foi visto na seção 5.2.1.3.

ORDEM DE PALAVRAS COM ADVÉRBIOS	
Advérbios	19
Outros constituintes com valor adverbial	2
TOTAL	**21**

Tabela 54 – Ordem de palavras com Advérbios

Primeiramente, serão observados dados com modificadores de predicado, tipicamente advérbios de modo, advérbios de localização temporal e espacial, e advérbios de quantidade e grau. Segundo Costa (2008), eles ocorrem, normalmente, após o verbo e não em posição pré-verbal. Além disso, a maior parte deles só ocorre na posição inicial da frase quando há contexto contrastivo (cf. seção 5). Em segundo lugar, serão analisados sucintamente os dados com advérbios modificadores de frase, como os advérbios conectivos e os avaliativos. Também segundo Costa (2008), eles ocorrem tipicamente em posição pré-verbal e podem ser inseridos no início da frase com facilidade. Para serem

posicionados entre o sujeito e o predicado, é necessária a entoação parentética, geralmente representada por vírgulas na escrita.

Os dados apresentados em seguida contêm verbos de localização temporal. Esses advérbios são especias, pois podem facilmente ocorrer na posição inicial e entre o sujeito e o predicado, sendo necessária uma entoação parentética neste último caso (cf. seção 5). Ao contrastar as propostas de tradução (293c)/(293d) e (294c)/(294d), é possível verificar que o problema não reside na ordem dos advérbios assinalados, mas sim em outros problemas de tradução presentes no texto: após a correção desses erros, é possível verificar que o advérbio "atualmente" e o constituinte com valor adverbial "do início" podem ser inseridos em diversas partes do texto sem prejudicar o sentido original do texto. Por isso, quanto à ordem das palavras, a tradução feita na empresa foi assinalada com a expressão "Possivelmente não natural" na presente pesquisa.

(293)

a. *by the same country that you currently reside in*
b. ? pelo mesmo país que **atualmente** está
c. pelo mesmo país em que você reside **atualmente**
d. pelo mesmo país em que você **atualmente** reside

(294)

a. *you should be able to start with the sign up process **from the scrach**.*
b. ? você deve começar o processo de se inscrever **do início**.
c. começar o processo de inscrição **do início**.
d. começar **do início** o processo de inscrição.

Já o exemplo seguinte, envolvendo um advérbio de grau, foi assinalado como contendo erro de ordem, pois o sentido do texto foi modificado. Esse tipo de advérbio ocorre tipicamente após o verbo quando exerce função de modificador de predicado. Todavia, há aqui dois verbos ("ajudar" e "navegar") e ao posicionar "muito" após o verbo "navegar" ele é interpretado como modificador do predicado que contém esse verbo, modificando o sentido do texto original. Por isso, a posição adequada para esse elemento é após o verbo "ajudar", como na proposta apresentada em (295c) em que também a palavra "navegar" foi alterada para outro termo mais adequado ao contexto.

(295)

a. *An attached screenshot of the issue can help us a lot navigating the problem as well.*
b. *Uma captura de tela anexa do problema pode nos ajudar a navegar **muito** o problema também.
c. Uma captura de tela anexa do problema pode nos ajudar **muito** a entender o problema também.

Como os outros modificadores de predicado, os advérbios de modo se posicionam tipicamente após o verbo, mas podem se mover dentro do sintagma: a posição de "totalmente" está correta em ambas as traduções da empresa (296b) e da proposta (296c).

(296)

a. *when you verify your account **fully** for getting access*
b. quando você verifica sua conta **totalmente** para obter acesso
c. quando você verifica **totalmente** sua conta para obter acesso

No caso da palavra "diretamente" apresentada a seguir, esse advérbio pode ser posicionado no final do SV, como nas traduções (297bi) e (297bii) encontradas nos dados fornecidas pela Unbabel. Porém, devido à distância entre o verbo e o advérbio nessas traduções, há outras opções mais naturais, como as posições propostas em (297d) e (297e):

(297)

a. *you to pay for the order by a (CARD BRAND)/(CARD BRAND) bank card **directly**.*
b.
    i.? pagar a encomenda através de um cartão bancário (CARD BRAND) / (CARD BRAND). **Diretamente**.
    ii.? a opção de pagamento através de um cartão bancário (CARD BRAND) / (CARD BRAND) **diretamente**.
c. *? pagar pela encomenda através de um cartão bancário (CARD BRAND) / (CARD BRAND) diretamente*
d. *pagar pela encomenda diretamente através de um cartão bancário (CARD BRAND) / (CARD BRAND)*
e. *pagar diretamente pela encomenda através de um cartão bancário (CARD BRAND) / (CARD BRAND)*

No caso do advérbio conectivo "também", a posição típica desse tipo de advérbio é pré-verbal, como na proposta (298c). Todavia, Costa (2008) ressalta que esse advérbio

também pode ocorrer em no fim da frase (298b), como apresentado na tradução feita pela empresa.

(298)

a. *An attached screenshot of the issue can help us a lot navigating the problem **as well**.*
b. Uma captura de tela anexa do problema pode nos ajudar muito a navegar o problema **também**.
c. Uma captura de tela anexa do problema **também** pode nos ajudar muito a entender o problema.

Os advérbios focalizadores podem modificar o predicado inteiro ou somente um elemento da frase (cf. seção 5). Esse tipo de advérbio deve ser inserido à esquerda do elemento que modifica. No caso da tradução (299b), fornecida pela empresa, o advérbio "apenas" foi inadequadamente repetido em duas partes do texto: no início do SV e antes do modificador. No texto original, esse advérbio tem escopo sobre o SP "*over the phone*". Para manter o sentido apresentado nesse texto, somente um dos advérbios pode ser mantido no texto de chegada, como apresentados nas propostas (299c) e (299d).

(299)

a. *As this is sensitive information we can provide this **only** over the phone.*
b. *Como esta é uma informação sensível, **apenas** podemos fornecer isso **apenas** telefone.
c. Como esta é uma informação sensível, **apenas** podemos fornecer isso por telefone.
d. Como esta é uma informação sensível, podemos fornecer isso **apenas** por telefone.

As locuçãoes adverbiais são sequências de mais de uma palavra que têm o mesmo comportamento dos advérbios. Esses elementos possuem uma organização interna tipicamente rígida que dificilmente pode ser alterada. Foi encontrado nos dados um caso que envolvia essa ordenação dentro de uma locução adverbial: a expressão "*once again*" foi traduzida por "uma vez mais" nos dados da empresa (300b). Apesar de a expressão "mais uma vez" (300c) ser mais comum em PB, a expressão utilizada em (300b) também é possível em PB e não é agramatical. Assim, a tradução (300b) proposta pela empresa foi assinalada com a expressão "Possivelmente não natural".

(300)

a. *send verification request **once again***
b. envie um pedido de confirmação **uma vez mais**
c. envie um pedido de confirmação **mais uma vez**

Considerando os dados tratados na presente subseção, somente um caso foi assinalado como efetivamente contendo erro de ordem devido à mudança no sentido (cf. 295b). Os outros 20 dados foram considerados possivelmente não naturais, pois havia várias possibilidades de ordenação e a ordem apresentada nos dados na empresa não estavam agramaticais.

Para finalizar, a tabela 55 abaixo apresenta o número de problemas de ordem no interior da frase segundo os três tipos de fenômenos encontrados: 4 dados foram assinalados como "ordem de palavras", pois há efetivamente erros de ordem, ou seja, constituem estruturas agramaticais; 23 dados não eram agramaticais, mas sim possivelmente não naturais, sendo, por vezes, possível encontrar outras propostas mais adequadas; 14 dados envolviam o fenômeno de variação em PB, no caso da presente subseção somente os complementos clíticos foram assinalados com essa expressão.

TIPOS DE PROBLEMAS DE ORDEM NO INTERIOR DA FRASE	
Ordem de Palavras	4
Possivelmente não natural	24
Variação em PB	14
TOTAL	**42**

Tabela 55 – Tipos de problemas de ordem no Interior da Frase

6.2.2.4 Síntese dos problemas de ordem encontrados nos dados

A tabela 56 a seguir apresenta um apanhado geral dos tipos de problemas encontrados nos dados envolvendo a ordem de palavras no interior do SN e no interior da frase, observados na seção 6.2.2. É possível verificar que aproximadamente 44.5% dos dados apresentavam efetivamente erros de ordem, levando à agramaticalidade das sentenças ou sintagmas nos quais estavam incluídos. Dentre esses dados errados, é de notar que certos casos possuem não somente erro de ordem, mas também outros tipos de erro como, por exemplo, concordância inadequada ou a falta de preposição. Em outros 28 dados, correspondendo a 36% dos casos, a ordem não era agramatical. No entanto, as posições desses constituintes eram interessantes para a discussão feita na presente pesquisa, pois foi possível encontrar outras posições mais naturais ou adequadas para os elementos anotados. Finalmente, 14 dados envolviam o fenômeno de variação no PB,

mais especificamente a posição dos clíticos na frase, fenômeno muito delicado, tendo em vista a coexistência entre a tendência proclítica do PB e presença de estruturas enclíticas nos textos escritos nessa variante.

PROBLEMAS DE ORDEM DE PALAVRAS NOS DADOS RECOLHIDOS	
Ordem de Palavras	15
Ordem de Palavras e outros erros	18
Possivelmente não natural	28
Variação em PB	14
TOTAL	**75**

Tabela 56 – Problemas de ordem de palavras nos dados recolhidos

Como já foi tratado na seção 2, a diminuição da influência da subjetividade é um dos grandes desafios na tradução automática e no processo de anotação. Tendo em vista as várias possibilidades de tradução da ordem na LC nos casos assinalados como "Possivelmente não natural" e "Variação em PB", pode ser difícil encontrar orientações ou regras gerais que auxiliem a correção dos anotadores e delimitem a anotação dos anotadores. Também fenômenos como a concordância por sentido, frequente em português, podem dificultar no aumento da objetividade do processo de anotação e correção das traduções feitas na empresa. A próxima seção 7 busca contribuir com a construção de algumas orientações para auxiliar no processo de anotação, tendo como base os dados observados ao longo da presente seção.

7. CONTRIBUTOS DO PRESENTE TRABALHO

A presente seção se dedica a fornecer orientações e sugestões gerais elaboradas a partir da análise do processo de anotação dos erros feito pelos anotadores da empresa, bem como dos principais problemas de concordância e ordem de palavras encontrados nos dados, já discutidos na seção 6. Esses contributos foram divididos em três partes: na subseção 7.1, serão apresentadas sugestões gerais com o objetivo de aprimorar duas das *Guidelines* elaboradas pela empresa já discutidas na presente pesquisa, nomeadamente as *Annotation Guidelines* e as *Linguistic Guidelines (PT-BR)*; na subseção 7.2, serão fornecidas sugestões para a avaliação e treinamento dos anotadores e editores da empresa da qual foram analisados os dados; a subseção 7.3 trata da possível criação de fóruns linguísticos.

7.1 Contributos para as *Guidelines*

O foco da presente seção é examinar algumas das questões que surgiram nas seções 6.1 e 6.2, dedicadas à análise dos dados fornecidos pela empresa, e fornecer sugestões para as *Annotation Guidelines* e as *Language Guidelines* que podem auxiliar na resolução de alguns dos problemas encontrados. Esta seção segue a seguinte organização: a subseção 7.1.1 apresenta uma reavaliação das orientações de segmentação e propõe algumas sugestões; a seção 7.1.2 contém sugestões gerais para o processo de categorização; e a seção 7.1.3 fornece sugestões específicas para as *Language Guidelines*, ressaltando os processos de anotação e de pós-edição.

7.1.1 Sugestões e reavaliação das orientações de segmentação

A partir da observação mais detalhada das orientações dadas pelas *Annotation Guidelines*, feita na subseção 6.1.1.1, foi possível notar a dificuldade em elaborar indicações de segmentação que possam incluir todos os casos de concordância e de ordem de palavras em PB, algumas vezes não ficando claro qual seria a maneira mais adequada de segmentar as unidades. É necessário lembrar que a empresa Unbabel traduz textos escritos de e para 28 línguas e possui uma tipologia de anotação bastante complexa, sendo difícil fornecer conselhos gerais de anotação que não sejam demasiado longos, mas que possam ser aplicados a todas essas línguas, algumas delas tipologicamente diferentes.

Tendo em conta esses aspectos, as divergências nas anotações são um fenômeno comum. No entanto, uma avaliação ideal da qualidade deve obter resultados o menos subjetivos possíveis, sendo um dos principais propósitos das *Annotation Guidelines* a diminuição da subjetividade entre anotadores humanos, para atingir assim uma maior concordância inter-anotadores.

As sugestões que se apresenta em seguida procuram respeitar os seguintes critérios: por um lado, as orientações devem fornecer instruções gerais que sejam flexíveis e possam ser utilizadas em diversos contextos linguísticos; por outro lado, elas devem ser específicas o suficiente para que os erros apresentados nos resultados sejam facilmente contabilizados e identificados. O foco das sugestões será a tentativa de solucionar as instruções de segmentação aparentemente ambíguas nas *Annotation Guidelines*, através da observação da variação na segmentação dos dados de *Agreement* e *Word Order* em PB feita por anotadores humanos.

Para auxiliar na elaboração das sugestões, na próxima subseção 7.1.1.1 serão revistas as ideias de Burchardt e Lommel (2014) acerca da dificuldade na segmentação de erros de *Word Order* e *Agreement* no *QTLaunchPad*. A escolha desses autores se deve ao fato de as suas orientações terem sido a base a partir da qual se desenvolveu a tipologia de anotação utilizadas na Unbabel (cf. seção 4). Em seguida, na subseção 7.1.1.2 serão fornecidas sugestões gerais e exemplos de segmentação para essas duas etiquetas, tendo em conta as ideias desses autores e as propriedades da ferramenta de anotação da Unbabel apresentada na seção 4.2.

7.1.1.1 Questões de segmentação

Durante a presente pesquisa, considerou-se proveitoso fazer uma conexão entre as ideias Burchardt e Lommel (2014) e as questões que surgiram durante a análise da segmentação feita na subseção 6.1, buscando assim solucionar certas dificuldades. Antes de discutir as etiquetas em foco, vale salientar que, segundo esses autores, é imprescindível a seleção do menor intervalo de texto possível, devendo ser selecionada somente a área necessária para a explicitação do problema, mas também devendo ser possível a seleção de mais de uma unidade quando necessário para especificar o erro (Burchardt e Lommel, 2014: 4). Nesse aspecto, as *Annotation Guidelines* dão orientações semelhantes, prescrevendo a seleção do menor intervalo para a correção do problema, apesar dos exemplos fornecidos pela empresa nem sempre seguirem essas orientações (cf. seção 6.1.1.1).

Em relação às instruções para segmentação de erros de *Agreement* nas *Annotation Guidelines* da empresa, foram três os principais casos que suscitaram dúvidas (cf. 6.1.1.1): a) a divergência entre a instrução de seleção do "*minimal spam to fix the issue*" e os exemplos de concordância verbal apresentados pela empresa, em que dois elementos foram selecionados, indo aparentemente contra o que foi indicado nessa instrução b) a ausência de exemplos de erros de concordância com outros traços como gênero, pessoa e caso; c) a escassez de exemplos que evidenciem a diferença na segmentação e na anotação de elementos pertencentes ao mesmo erro e à mesma unidade sintática e de elementos com o mesmo tipo de erro, mas em unidade sintáticas distintas.

As *Guidelines* do *QTLaunchPad* ressaltam os desafios na segmentação de erros de concordância, pois as porções discordantes podem estar separadas no texto. Nesses tipos de casos, Burchardt e Lommel (2014: 5) providenciam duas instruções: a) "if two items disagree and it is readily apparent which should be fixed, mark only the portion that needs to be fixed"; b) "if two items disagree and it is not clear which portion is incorrect, mark both items and mark them for Agreement". Na ferramenta que segue as instruções do *QTLaunchPad* são selecionadas somente as unidades que estão incorretas, sendo aceita também a seleção de mais unidades nos casos em que há dúvidas acerca de qual dos elementos está originalmente incorreto. Os seguintes excertos, retirados do texto desses autores exemplificam esses casos:

(301)

a. *Incorrect markup*: The **man and its** companion were business partners.
b. *Correct minimal markup*: The man and **its** companion were business partners.

(302)

a. *Incorrect markup*: The man whom they saw on Friday night at the store **were** very big.
b. *Correct minimal markup*: The **man** whom they saw on Friday night at the store **were** very big.

(Burchardt e Lommel, 2014: 5)

Segundo esses autores, em (301) está claro que o elemento discordante é o pronome "*its*", pois ele se refere a "*the man*" e deveria apresentar a forma "*his*". Por isso, respeitando o princípio de seleção mínima, somente a unidade [its] deve ser selecionada, como em (301b). Já em (302), dado que o único contexto fornecido é a frase apresentada, não fica claro se o erro de concordância está no verbo ou no sujeito. Nesse tipo de caso,

os autores afirmam que devem ser selecionadas as duas unidades: o núcleo do SN [man] e o verbo [were], como em (302b).

É possível observar que o exemplo (302) apresentado pelos autores é muito semelhante ao exemplo (222a) das *Annotation Guidelines* da Unbabel, em que também foram selecionadas duas unidades: "*The **man** whom they saw on Friday **were** very big*". Essa falta de clareza acerca de qual das unidades apresenta erro de concordância e a necessidade de marcar ambas as unidades sob a etiqueta *Agreement* geralmente não se aplica aos erros anotados com a ferramenta da Unbabel, pois, em primeiro lugar, também é disponibilizado aos anotadores o texto de partida, possibilitando a resolução de ambiguidades e a confirmação dos traços de concordância no texto de partida. Em segundo lugar, como já se referiu sucintamente na nota 29 da seção 6.1.2 e na Tabela 24 da seção 6.1.2.2.1 (cf. Tabelas 3.1 e 3.2 no Anexo 3), a Unbabel faz a distinção entre as categorias *Spelling* e *Agreement*: por um lado, se o sujeito "*the man whom they saw on Friday*" estava no plural no texto original, mas foi traduzido no singular, há erro de *Accuracy* e a etiqueta *Spelling* deveria ter sido aplicada a esse sujeito; por outro lado, se o texto apresenta uma discordância de traços, como ocorre claramente ao se observar a relação entre o sujeito "*the man whom they saw on Friday*" e o verbo "*were*", mesmo sem observar o texto de partida, há erro de *Fluency* e a etiqueta *Agreement* deve ser a escolhida para a anotação no verbo. No caso do exemplo citado, há um erro de concordância verbal, sendo o sujeito da oração o controlador dos traços de pessoa e número do verbo. Por isso, este verbo é a unidade que deveria ter sido selecionada com a etiqueta *Agreement*.

Tendo em conta a disponibilização do texto de partida, bem como as diferenças entre as etiquetas da Unbabel citadas no parágrafo anterior, e aplicando-se o princípio de seleção mínima, citado quer pelas *Annotation Guidelines* quer pelas *Guidelines* da *QTLaunchPad*, e a noção de controlador dos traços de concordância, pode-se afirmar que selecionar somente a unidade em que se encontra o erro seria possivelmente o procedimento mais adequado para segmentar os casos que envolvem a etiqueta *Agreement*. Além disso, a aplicação dessa instrução poderia evitar também a anotação de unidades que foram corretamente traduzidas.

Seguindo esta linha de raciocínio, nos dados apresentados na tabela 57, a segmentação feita pelos anotadores em A94 é adequada, pois somente o verbo [podem] foi selecionado; já em A33, sugere-se que somente a unidade que inclui a preposição e o determinante, [do], seja selecionada. Também seguindo esse princípio, nas sugestões apresentadas não foi necessário selecionar como erro os núcleos [maioria] e [mensagem],

que determinam os traços morfossintáticos dos elementos problemáticos e que se encontram ambos corretamente traduzidos quando se compara o texto de partida e o texto de chegada.

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Sugestões de segmentação**
A94	*Most visual issues can be resolved..*	*A maioria dos problemas visuais [podem] ser resolvidos...*	podem	podem
A29	*Full page screenshot of the error message..*	*Página completa imagem de captura de tela [do] [mensagem] de erro...*	do/mensagem	do

Tabela 57 – Exemplos de sugestões de segmentação (*Agreement*)

Nos casos de seleção de *Word Order*, as seguintes informações das *Guidelines* levantaram dúvidas (cf. seção 6.1.1.1): a) a falta da apresentação de definições e de explicações acerca das diferenças entre palavras contíguas, descontínuas e adjacentes e da importância desses aspectos no processo de anotação de *Word Order*; b) a discrepância entre a instrução que recomenda a seleção da "*shortest portion of text that could be moved to solve the problem*" e os exemplos que apresentam a seleção de mais de uma palavra, mesmo sendo possível mover somente uma das palavras para que o erro de ordem seja resolvido, indo aparentemente contra aquilo que foi indicado nessa instrução; c) a falta de exemplos de segmentação e anotação de erros de *Word Order* em que a palavra mal posicionada não se encontra na posição incorreta em relação a outra palavra, mas sim em relação a um extenso bloco de palavras, como ocorre frequentemente no caso de advérbios ou na ordem relativa entre complementos verbais, por exemplo.

No caso dos erros de ordem de palavras, também Burchardt e Lommel (2014: 4-5) ressaltam a dificuldade na segmentação, pois esse tipo de erro pode envolver longas partes do texto. Por isso, os autores recomendam a seleção da "shortest portion of text (in number of words) that could be moved to fix the problem. If two portions of the text could resolve the problem and are equal in length, mark the one that occurs first in the text" (Burchardt e Lommel, 2014: 4). Os seguintes exemplos dos autores ilustram as

formas corretas e incorretas de segmentar os erros de *Word Order*, segundo as suas ideias:

(303)

a. *Incorrect markup*: The **telescope big** observed the operation.
b. *Correct minimal markup*: The **telescope** big observed the operation.

(304)

a. *Incorrect markup*: The eruption **by many instruments was recorded**.
b. *Correct minimal markup*: The eruption by many instruments **was recorded**.

(305)

a. *Incorrect markup*: The **given policy in the manual user** states that this action voids the warranty.
b. *Correct minimal markup*: The **given** policy in the **manual** user states that this action voids the warranty.

(Burchardt e Lommel, 2014: 4-5)

No exemplo (303a), ambos os termos "*telescope*" e "*big*" devem trocar de posição, dado que, neste contexto, o inglês só admite a ordem adjetivo-nome, sendo este um caso em que o problema de ordem é interno ao SN. Seguindo o princípio da seleção mínima e a instrução de selecionar o primeiro termo incorreto que ocorre no texto, os autores citados afirmam que deve ser selecionada a unidade [telescope]. No exemplo (304a), há um erro de ordem ao nível frásico: o agente da passiva se encontra entre o sujeito e o SV de que faz parte. Nesse caso, apesar de "*by many instruments*" ter ocorrido primeiro na frase, o deslocamento de [was recorded], porção de texto menor, resolveria o problema, por isso esse bloco seria o segmento a ser anotado. Em (305a), os autores apresentam dois erros de ordem diferentes: por um lado, a forma participial "*given*" convertida em adjetivo deve ocorrer à direita do nome que modifica (*policy*), uma vez que está seguida de um complemento (*the user's manual*); por outro lado, esse complemento citado contém um erro, pois em inglês o nome que refere o possuidor, "*user"*, deve ocorrer antes do nome que refere o objeto possuído, "*manual*" (*the user's manual*). Nesse caso, os autores solicitam que os anotadores selecionem as unidades [given] e [manual] separadamente, com duas etiquetas *Word Order*, pois há dois problemas diferentes.

Como já foi mencionado na subseção anterior, também as *Annotation Guidelines* da empresa prescrevem a seleção da menor porção de texto que poderia ser movida para resolver o problema e recomendam a seleção e a anotação individual das unidades com erros diferentes pertencentes à mesma categoria de erro. Porém, como é possível

observar nos excertos (223), (224) e (225), na seção 6.1.2.1, sempre mais de uma unidade é selecionada nos exemplos da empresa. A aplicação das propostas de anotação de *Word Order* defendidas por Burchardt e Lommel (2014) poderia resolver certas incongruências encontradas nas *Annotation Guidelines* e auxiliar os anotadores quando estes possuem dúvidas acerca da unidade a anotar quando as unidades são muito extensas ou quando o erro envolve duas unidades do mesmo tamanho. Isso poderia evitar a seleção de unidades que não se encontram na posição errada e diminuir a já citada tendência dos anotadores para selecionar o elemento à esquerda ou à direita da posição em que a unidade deveria estar (cf.seção 6.1.1).

A tabela 58 apresenta alguns exemplos de segmentação de erros de ordem de palavras segundo as ideias dos autores citados nos parágrafos anteriores. Em W76, os termos "cobrança" e "ingresso" estão na posição incorreta, pois a ordem mais adequada na tradução seria "*um ingresso de cobrança*". Nesse caso, a sugestão seria somente a seleção de [cobrança], pois é a primeira palavra que ocorre no texto de chegada e ambas as palavras "cobrança" e "ingresso" possuem o mesmo tamanho. Em W54, a tradução mais adequada segundo o contexto seria "*nossos produtos DJ*", pois o termo "DJ" é o nome próprio de um dos produtos da empresa. Somente a palavra [produtos] deveria ter sido selecionada nesse caso, pois é a primeira palavra na posição incorreta e o seu deslocamento entre os termos "nosso" e "DJ" resolveria o problema de ordem, apesar de ainda ser necessário corrigir o problema de concordância e adicionar a preposição e o determinante.

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Sugestões de segmentação**
W76	*Please send them a billing ticket through URL-0...*	*Por favor enviar eles um **cobrança** ingresso através de URL-0...*	cobrança/ingresso	cobrança
W54	*It is possible to purchase our DJ products directly...*	*É possível comprar **produtos** do nosso DJ diretamente...*	produtos/nosso/DJ	produtos

Tabela 58 – Exemplos de dados adequadamente segmentados (*Word Order*)

Finalmente, nas instruções gerais para a segmentação de todas as etiquetas, as *Annotation Guidelines* fornecem duas orientações relevantes para presente pesquisa: se há duas ou mais unidades repartidas no texto que formam um único erro, o anotador deve selecionar todas as unidades e categorizá-las na mesma etiqueta; mas se há dois ou mais erros individuais que pertencem ao mesmo tipo de erro, as unidades devem ser selecionadas e categorizadas individualmente. Entretanto, não há uma definição clara ou exemplos que ajudem diferenciar entre as unidades que fazem parte do mesmo erro e as unidades que fazem parte de erros distintos. Para exemplificar essa questão, no excerto (306) a seguir "deve" e "superior" são o mesmo erro pois têm o mesmo controlador para cada uma das concordâncias envolvidas? Ou são erros distintos pois fazem parte de unidades sintáticas distintas e devem ser selecionados em etiquetas separadas?

(306)

a. *All documents must be older than 3 months*.
b. Todos os documentos **deve** ser **superior** a 3 meses.

Essa é uma questão que, ao ser solucionada, poderia auxiliar a anotação feita na empresa, pois esse tipo de dúvida pode levar à anotação não-uniforme dos dados. Na segmentação feita nos próximos parágrafos da presente pesquisa, procurou-se encontrar um meio termo entre essas duas alternativas. Porém, ainda não encontramos uma solução infalível para essa questão. Logo, esse pode ser o tema de trabalhos futuros feitos na empresa. Por ora, sugere-se a inserção da noção de unidade sintática e de controlador da concordância em ambas as *Guidelines* da empresa, pois esses temas podem auxiliar os anotadores a fazerem uma reflexão sobre as diferentes relações gramaticais presentes nos textos de partida e chegada.

7.1.1.2 Sugestões para a segmentação

Na presente subseção, apresenta-se sugestões para as *Annotation Guidelines* que poderiam auxiliar na uniformização da segmentação dos dados relacionados com as etiquetas de *Agreement* (em 7.1.1.2.1) *e Word Order* (em 7.1.1.2.2). Quanto ao seu conteúdo, essas subseções seguem a seguinte organização: são listadas as principais sugestões para a segmentação de erros relativos a etiqueta em foco; logo em seguida, a aplicação dessas sugestões é apresentada através da reformulação de regras e de exemplos elaborados para os propósitos da presente subseção.

7.1.1.2.1 Sugestões para a segmentação de *Agreement*

Apresenta-se a seguir propostas para a reformulação da segmentação de elementos com erro de *Agreement*. Durante a presente pesquisa, foram elaborados exemplos da aplicação das sugestões apresentadas, disponíveis na Tabela 7.1 do Anexo 7 e citadas nas notas de rodapé. As sugestões relacionadas com as *Language Guidelines* serão abordadas na seção 7.1.3, que trata desse aspecto.

(S1) Reformulação das regras de segmentação de *Agreement*, baseando-se nas instruções já presentes nas *Annotation Guidelines* e nas ideias de Burchardt e Lommel (2014) que poderiam ser aplicadas ao contexto da ferramenta de anotação da empresa.

(S2) Introdução sucinta da noção de controlador da concordância e da existência de diferentes unidades sintáticas nas *Annotation Guidelines*, para explicitar em qual elemento efetivamente se encontra a discordância de traços. Inclusão desses conceitos de maneira mais detalhada também nas *Language Guidelines* com exemplos específicos da língua tratada.

(S3) Adaptação dos exemplos de segmentação já apresentados nas *Annotation Guidelines* para que estejam de acordo com a reformulação das regras de segmentação (caso essa sugestão seja considerada válida para o contexto da empresa) e inclusão de uma maior variedade de tipos de erros de concordância através da inserção de exemplos:

a) envolvendo cada um dos quatro traços morfossintáticos ligados à concordância (gênero, número, pessoa e caso[35])[36];

b) cujos elementos incorretos pertencem a diferentes categorias gramaticais, inseridos numa maior variedade de sintagmas e estruturas gramaticais (não envolvendo somente a concordância verbal)[37];

c) que demonstrem a diferença de segmentação de concordância para elementos inseridos na mesma unidade sintática (os elementos devem ser anotados juntos sob a mesma etiqueta) e elementos inseridos em unidades

[35] Não serão apresentados na presente pesquisa exemplos com o traço de "caso", pois não se aplica ao PB.
[36] Cf. na Tabela 7.1 do Anexo 7: EA1 (número), EA3 (gênero), EA6 (pessoa).
[37] Cf. na Tabela 7.1 do Anexo 7: EA1 (verbo), EA3 (artigo indefinido e adjetivo), EA4 (demonstrativo e artigo definido) e EA5 (clítico).

sintáticas distintas (os elementos devem ser anotados separadamente em duas etiquetas distintas)[38].

Considerando o que foi apresentado nas sugestões, sugere-se a inserção da seguinte orientação para a segmentação dos casos envolvendo a etiqueta *Agreement*.

> "Também na segmentação dos erros de concordância, mantenha a marcação mínima dos erros: se duas unidades estão em discordância, selecione somente a unidade que efetivamente contém o erro de concordância. Lembre-se de que o núcleo do SN e o sujeito da oração, entre outros, são frequentemente os controladores dos traços morfossintáticos (cf. *Language Guidelines*), por isso esses elementos geralmente não apresentam erro de *Agreement*, mas sim outro tipo de erro que se encaixa em outra etiqueta, como, por exemplo, *Spelling,* sendo sempre necessário verificar o texto de partida.
>
> Nos casos em que os elementos com erro de concordância formam um único erro, ou seja, pertencem a mesma unidade sintática (cf. *Language Guidelines*), selecione cada um desses elementos e categorize-os juntos na mesma etiqueta. Nos casos em que há vários erros do mesmo tipo, mas pertencentes a unidades sintáticas distintas, selecione e categorize cada um dos elementos separadamente."

7.1.1.2.2 Sugestões para a segmentação de *Word Order*

Nesta subseção são apresentadas algumas sugestões que podem auxiliar na uniformização da segmentação dos dados com erros de *Word Order.* Como na subseção anterior, os exemplos nas notas de rodapé se referem aos excertos apresentados na tabela 7.2 do Anexo 7.

(S4) Reformulação das orientações presentes nas *Annotation Guidelines* para a segmentação de erros de *Word Order*, baseando-se nas ideias de Burchardt e Lommel (2014) que poderiam ser úteis no contexto de anotação feito na empresa.

(S5) As *Annotation Guidelines* mencionam a existência de palavras contíguas, descontínuas e adjacentes. Porém, durante a presente pesquisa, considerou-se mais apropriado focar na noção de unidade sintática, diferenciando entre palavras vizinhas que fazem parte do mesmo constituintes e palavras vizinhas que pertencem a constituintes

[38] Comparar EA3 (determinante e modificador no mesmo SN) e EA4 (dois determinantes em SN distintos), presentes na Tabela 7.1 do Anexo 7.

distintos. Essa noção deve também ser incluída de maneira mais detalhada nas *Language Guidelines*.

(S6) Reformulação dos exemplos de segmentação de erros de *Word Order* já presentes nas *Annotation Guidelines*, considerando-se as sugestões anteriores e inclusão de uma maior variedade de tipos de erros de ordem de palavras, através da inserção de exemplos de segmentação em que os erros de ordem:

a) envolvem contextos em que a unidade de segmentação está mal posicionada em relação a uma unidade sintática inteira, como no caso da ordem de complementos verbais, sujeitos frásicos ou sintagmas adverbiais[39];

b) não podem ser resolvidos através da troca de lugar entre dois elementos vizinhos[40];

c) envolvem elementos vizinhos pertencentes ao mesmo constituinte e elementos vizinhos pertencentes a constituintes distintos[41];

d) explicitem a diferença entre elementos que envolvem o mesmo erro de ordem e devem ser anotados juntos na mesma etiqueta e elementos que devem ser anotados separadamente, pois envolvem erros de ordem distintos[42].

Tendo em vista as sugestões expostas acima acerca da segmentação de erros relativos à *Word Order*, a inserção da seguinte orientação é sugerida:

> "No caso dos erros de *Word Order*, respeite o princípio de marcação mínima: selecione a menor porção de texto que poderia resolver o erro de ordem ao ser movida. Caso o deslocamento de uma ou de outra unidade de mesma extensão, em número de palavras, possa resolver o problema de ordem, selecione somente a primeira que ocorre no texto.
>
> Repare que há palavras vizinhas que fazem parte do mesmo constituintes e palavras vizinhas que pertencem a constituintes distintos. Ao segmentar e anotar as unidades mal posicionadas, é importante verificar as diferentes relações sintáticas presentes no texto para não anotar juntos mesma etiqueta de *Word Order* elementos que não fazem parte do mesmo erro."

[39] Cf. na Tabela 7.2 do Anexo 7: o advérbio em EW5.
[40] Cf. na Tabela 7.2 do Anexo 7: EW7 e EW6.
[41] Comparar EW2 (advérbio e determinante + núcleo do SN) e EW4 (complemento e núcleo do SN), inseridos na Tabela 7.2 do Anexo 7.
[42] Comparar EW5 (palavras pertencentes a unidades sintáticas distintas, anotadas separadamente) e EW6 (palavras pertencentes a mesma unidade, anotadas juntas), inseridos na Tabela 7.2 do Anexo 7.

7.1.2 Sugestões para a categorização

As principais confusões feitas pelos anotadores da Unbabel durante a categorização de erros de *Agreement* e *Word Order* já foram discutidas na seção 6.1.2. Na presente seção, relembra-se alguns pontos mais controversos, que envolvem estruturas que não são agramaticais em PB, mas também não são completamente adequadas no contexto de chegada, pois são menos naturais ou correspondem a áreas de variação no PB. Esses casos são mais difíceis de categorizar, pois geralmente envolvem a subjetividade do editor e do anotador. Na presente pesquisa, casos desse tipo não foram categorizados como erro, mas sim assinalados com as expressões "Possivelmente não natural", "Envolvendo concordância semântica" e "Variação em PB" (cf. seção 6.2). Através da descrição apresentada na seção 5 e da análise feita na seção 6, foi possível verificar a dificuldade em elaborar regras gerais que possam dar conta desses casos.

Por um lado, a empresa, através das *Annotation Guidelines* e das *Language Guidelines*, ressalta a importância de o texto soar natural na LC e solicita que os editores e anotadores não se prendam à estrutura do texto na LP. Na ferramenta de anotação há um sistema de classificação da fluência do texto traduzido (cf. seção 3) e na tipologia de erros há a etiqueta "*Overly Literal*", dedicada aos casos em que o texto de chegada se limita demasiadamente à estrutura do original, causando problemas de interpretação. Esses pontos demonstram a preocupação da empresa com a naturalidade dos textos traduzidos.

Por outro lado, não há instruções claras sobre como anotar os já citados casos que não são agramaticais e não causam problemas de interpretação, mas não soam muito naturais na LC. Quando pensamos sobre esses casos, um leque de questões e possibilidades para a sua anotação se abre: tendo em vista que não se trata de sequências efetivamente agramaticais, é mais adequado continuar a utilizar somente a classificação de "*Fluency*" para avaliar a naturalidade do texto? Seria o caso de criar mais um tipo de categoria "*Not natural in Target Language*"? Ou seria melhor utilizar a etiqueta "*Overly Literal"* e assinalar esse tipo de dado com a severidade "*minor*"?

Tendo em vista a extensão do presente trabalho, responder a essas questões não entrou na lista dos objetivos da pesquisa. Mas aproveitamos para citar essas dificuldades, que podem ser objetos de trabalhos futuros que verifiquem a influência desses aspectos na nota final de tradução e o melhor procedimento para anotar esses casos não agramaticais, mas com impacto na fluência do texto e sujeitos à opinião subjetiva do anotador e do editor. Quanto às sugestões, acredita-se que a empresa deveria mencionar

a existência desses tipos de caso nas suas duas *Guidelines* através de exemplos para esclarecer os anotadores e os editores sobre as diferenças entre sequências agramaticais e sequências não-naturais.

Para o aprimoramento da categorização feita na empresa, a presente pesquisa sugere também o desenvolvimento de uma *Decision Tree* (DT), ferramenta que facilita o processo de decisão dos anotadores através de questões que podem ser respondidas por "sim" ou "não". A pesquisa de Figueira (2018) já apresenta uma versão detalhada de uma *Decision Tree* que auxiliou a anotação feita por dois anotadores diferentes, tendo o inglês como LP e o alemão como LC. A partir dos resultados dessa análise, o autor concluiu que "the annotation accuracy is apparently higher when an annotator uses a DT, especially in the categorization of difficult errors" (Figueira, 2018: 41). Como esse autor, a presente pesquisa também acredita que a elaboração de DTs poderia auxiliar no processo de anotação da Unbabel, sendo essa a principal sugestão para a melhoria dos resultados de categorização dos anotadores.

7.1.3 Sugestões para as *Language Guidelines*

As sugestões apresentadas a seguir foram feitas a partir da análise dos erros de concordância e ordem de palavras encontrados dados fornecidos pela empresa, feita na seção 6.2. Vale salientar que um dos objetivos da seção 5, destinada à descrição das estruturas, foi fornecer uma descrição dos fenômenos de concordância e de ordem de palavras que possam auxiliar na elaboração dessas orientações em PB. Também a seção 6 ressalta algumas diferenças entre o inglês e o PB e apresenta uma análise mais detalhada de certos erros que pode ser utilizada na elaboração dessas orientações. Como já foi visto na seção 3.3, as *Language Guidelines* de PTBR já apontam sucintamente para a importância de verificar a ordem das palavras no texto de chegada e indicam uma das diferenças entre a ordem dos clíticos em PB e PE. Tendo em vista serem casos muito problemáticos em PB, consideramos que a empresa poderia ampliar esses temas.

No caso da ordem dos clíticos e dos advérbios, é impossível fazer regras gerais que possam dar conta de todos os contextos em que esses elementos ocorrem em PB. Também no caso da concordância semântica e da concordância com infinitivos flexionados pode ser difícil orientar os anotadores e os pós-editores, pois nesses casos há mais de uma possibilidade de tradução para o excerto. Sabendo-se que são áreas críticas do PB, as sugestões apresentadas pretendem se certificar de que os editores e os anotadores estão familiarizados com esses fenômenos. Nesses casos, consideramos

importante que as *Language Guidelines* indiquem as melhores técnicas para a tradução e pós-edição desses tipos de estruturas.

As sugestões apresentadas a seguir possuem a seguinte organização: em (S7) serão fornecidas sugestões gerais para as *Language Guidelines* de todas as línguas traduzidas pela empresa; a sugestões em (S8) e (S9) indicam orientações para as *Language Guidelines* específicas do PB e tratam da concordância e da ordem de palavras, respectivamente. Algumas das propostas apresentadas em (S7) já foram mencionadas anteriormente nas seções 7.1.1 e 7.1.2. Em (S8) e (S9), os exemplos mencionados nas notas de rodapé já foram explorados na seção 6.2 da presente pesquisa.

(S7) Apresentar, de maneira geral e sucinta, as seguintes noções gramaticais básicas nas *Language Guidelines* de todas as línguas trabalhadas na Unbabel:

a) as diferentes unidades sintáticas que compõem as frases.

b) o papel do controlador da concordância.

c) as diferenças entre traduções agramaticais e traduções pouco naturais.

d) as diferenças entre excertos com erro de fluência gramatical (*Fluency*) e excertos com erro de fidelidade ao texto original (*Accuracy*).

(S8) Inserir nas *Language Guidelines* específicas do PB, os seguintes aspectos sobre o funcionamento da concordância em PB:

a) a pós-edição da concordância no caso de estrangeirismos[43];

b) os casos em que há simultaneamente erro de ordem e erro de concordância no mesmo elemento[44];

c) o uso do símbolo "()" no caso da alternância de traços em PB[45];

d) a importância de identificar as diferentes relações sintáticas para estabelecer a concordância[46];

e) a existência do fenômeno de concordância semântica em PB[47];

f) a importância de verificar o referente dos sujeitos nulos para avaliar a gramaticalidade da concordância[48];

[43] Cf. (249) "*__um__ *Venture Builder*".
[44] Cf. (233) "*__pré-pago__ instituições financeiras".
[45] Cf. (217) "*purchase__/s__*", "*a compra__/s__" e "a__(s)__ compra__(s)__".
[46] Cf. (235) "sua __caixa de entrada__ de e-mail __registrada__" ou "sua caixa de entrada __do e-mail registrado__".
[47] Cf. (248) "A maioria d__os problemas__ visuais __podem__ ser..." e "A __maioria__ dos problemas visuais __pode__ ser...".
[48] Cf. (306) "*__Todos os documentos__ devem: Mostrar a data de emissão e [-] não __deve__ ser superior a 3 meses".

g) a importância de verificar a referência de clíticos para avaliar a gramaticalidade da concordância[49];

h) a possibilidade de escolher o não-flexionado ou o infinitivo flexionado em certos contextos[50].

(S9) Inserir nas *Language Guidelines* específicas do PB, os seguintes aspectos sobre o funcionamento da ordem de palavras nesta variedade:

a) a importância de conhecer as diferentes relações sintáticas para traduzir o sentido do texto original usando a ordem de palavras adequada[51];

b) durante a tradução do SN, por vezes inserir preposição antes do nome (com função de modificador ou complemento do SN) ou mudar a sua categoria gramatical (caso seja possível), pois nomes não podem modificar diretamente o núcleo do SN[52];

c) as diferenças semânticas entre os adjetivos que podem ser posicionados antes dos núcleos do SN e os que não podem[53];

d) a possibilidade de os constituintes frásicos poderem ocupar diversas posições na frase. Fazer comparações entre posições que: a) são pouco naturais, mas gramaticais; b) causam mudanças no sentido do texto original; ou c) são agramaticais[54];

e) as possíveis posições dos clíticos em PB, em que há posições: a) agramaticais, b) gramaticais mais naturais em PB; e c) gramaticais pouco naturais[55]. Fazer uma discussão acerca da escolha, durante o processo de pós-edição, entre as regras tradicionais e as regras efetivamente usadas na escrita de falantes cultos em PB;

f) as várias posições que os advérbios podem ocupar em PB. Fazer um paralelo entre posições: a) gramaticais que alteram o sentido original; b)

[49] Cf. (254) "Terei muito prazer em ajudá-**lo**" e "Terei muito prazer em ajudá-**la**".
[50] Cf. (283) "para ajudar viajantes a **programar** as suas viagens e encontrar..." e "para ajudar viajantes a **programarem** as suas viagens e encontrarem...".
[51] Cf. (275) "*a billing **ticket***", "uma cobrança de **ingresso**" e "um **ingresso** de cobrança".
[52] Cf. (280a) "*to **bank** inquiries*", (280c) "aos inquéritos **bancários**" e (280d) "aos inquéritos **do banco**".
[53] Cf. (270b) "***secreto** painéis" e (270c) "um **novo** plano de assinatura".
[54] Cf., respectivamente, a) (284b) "você indicou no questionário **o seguinte**"; b) "você ajudou **o viajante**" e "**o viajante** ajudou você"; e c) (244b) "*para ajudar a **viajantes** programar as suas viagens".
[55] Cf., respectivamente, a) (291d) "*Para que eu possa **a** ajudar", b) "Para que eu possa ajudá-**la**" e c) (292b)a"??para **o** ajudar a escolher".

gramaticais que mantêm o sentido do original; c) gramaticais pouco naturais[56]; e d) agramaticais[57].

7.2 Sugestões de avaliação e treinamento

Como já foi explorado nas seções 3 e 4 a avaliação é um processo muito importante na Unbabel. O sistema de tradução automática e os pós-editores são constantemente avaliados para que a tradução final fornecida pela empresa tenha níveis adequados de qualidade. As sugestões da presente seção buscam auxiliar a avaliação e o treinamento dos anotadores e pós-editores do par linguístico inglês-PB, baseando-se na análise feita nas seções 6.1 e 6.2.

Quanto às sugestões para a etapa de avaliação, nas próximas subseções é revisto sucintamente as duas sugestões principais: os testes com respostas múltiplas (7.2.1) e o *Golden Text* (7.2.2). A partir dessas avaliações, a empresa pode efetuar um treinamento de pontos já identificados como problemáticos, comparando, através das alternativas e textos propostos, categorias aparentemente ambíguas ou fenômenos do PB que causam dúvidas.

Para a etapa de treinamento, se sugere que as avaliações possuam um sistema de *feedback* automático, através do qual as pessoas avaliadas, caso tenha respondido incorretamente, recebam uma mensagem, no final de cada pergunta, com instruções e informações acerca do problema ou fenômeno abordado no teste. Através desse método, o anotador e o pós-editor podem aprender de imediato através de seus próprios erros.

Os excertos presentes nos exemplos de avaliações a seguir foram retirados e adaptados a partir dos dados fornecidos pela Unbabel. Nas alternativas dos testes e nos textos de cada *Golden Text*, foram inseridas algumas dificuldades com base nas confusões encontradas nos dados fornecidos (cf. seções 6.1 e 6.2). As propostas de soluções e mensagens de treinamento inseridas nas avaliações foram feitas segundo as sugestões (S1) à (S9) apresentadas na seção anterior. Para mais exemplos de testes e *Golden Texts* feitos durante a presente pesquisa, conferir os Anexos 8 e 9[58], respectivamente.

---

[56] Cf., respectivamente, a) (295) "*can help us a lot navigating the problem*" / "pode nos ajudar a navegar **muito** o problema"; b) (296b) "você verifica sua conta **totalmente**"/(296c) "você verifica **totalmente** sua conta"; c) "??você **totalmente** verifica sua conta".

[57] Cf. (148b) "*...podem **muito** aparecer".

[58] A partir das sugestões apresentadas nos Anexos 8 e 9, é possível construir outros tipos de avaliações, como perguntas com respostas abertas ou testes de escolha entre as opções "verdadeiro" ou "falso". Porém, por motivos de espaço, não foi possível apresentar mais exemplos com esses tipos de teste, que podem ser objeto de trabalhos futuros.

7.2.1 Sugestões de testes de múltipla escolha

Alguns exemplos de testes de múltiplas escolhas foram inseridos nas tabelas 59 e 60 apresentadas a seguir. Quanto à organização das tabelas, o código apresentado na primeira coluna se refere ao número do teste na lista de exemplos apresentadas no Anexo 8. Na segunda coluna da tabela, foram inseridos os textos dos testes com a questão e as alternativas. Na terceira coluna, há uma mensagem de treinamento, a ser revelada ao anotador avaliado nos casos em que foi escolhida uma alternativa inadequada. Note que as resoluções propostas foram feitas com base nas sugestões apresentadas na seção 7.1.

O foco dos testes na tabela 59 é o sistema de segmentação (T1) e a categorização (T14). Na tabela 60 são trabalhados os conhecimentos dos anotadores e editores acerca da tradução da ordem e concordância em PB.

Como já foi discutido na descrição da concordância e ordem de palavras em PB (seção 5) e na análise dos erros ligados à aspectos linguísticos (seção 6.2), é possível fazer uma divisão geral entre os casos que possuem um funcionamento mais constante, sendo possível identificar regras fixas (como, por exemplo, a concordância dentro do SN), e os casos que estão sujeitos aos fenômenos mais inconstantes, sendo difícil estabelecer regras gerais acerca do seu funcionamento (como a variação da concordância em PB e a posição dos elementos na frase). Nesse último tipo de caso, geralmente há mais de uma possibilidade correta de traduzir em PB, fazendo com que a subjetividade do anotador e do pós-editor influenciem no resultado final da tradução. Tendo em vista esses aspectos, considerou-se interessante inserir testes em que há mais de uma alternativa correta devido às diferentes possibilidades de traduzir o excerto apresentado em PB. Esse é o caso do teste T25 (e de outros testes apresentados na Tabela 8.7 do Anexo 8).

Cód.	Texto do teste	Mensagem de treinamento
T1	**A partir da observação do texto de partida, escolha a melhor alternativa de segmentação do(s) erro(s) presente(s) na tradução em PB:** - *Screenshots of the error message* - Imagens de captura de tela do mensagem de erro. a) [do mensagem] **b) [do]** c) [do] [mensagem] d) [o]	**b) Sugestão de resolução** Incluir somente a unidade incorreta. A unidade mínima de seleção é a palavra inteira. Considerando-se que o determinante "o" está unido à preposição "de", toda a unidade [do] deve ser selecionada. Nesse caso, "mensagem" é o controlador da concordância e não foi mal traduzido, logo não deve ser selecionado.
T14	**A partir da observação do texto de partida, escolha a melhor alternativa de categorização do(s) erro(s) presente(s) na tradução em PB:** -*Do you have a subscription plan?* -Você possui um [**assinatura**] plano? a) Word Order + Wrong Preposition **b) Word Order + Omitted Preposition** c) Word Order d) Overly Literal	**b) Sugestão de resolução** A tradução correta da unidade sintática na qual se encontra a unidade selecionada é "um plano de assinatura". Apesar da influência da ordem do texto de partida no texto de chegada, levando a uma tradução muito literal, prefira categorizar através de etiquetas mais específicas. Há erro de ordem, pois "assinatura" deve estar após o seu núcleo "plano". Não confunda as etiquetas *Wrong Preposition* e *Omitted Preposition*: o primeiro se refere ao uso incorreto de uma preposição; o segundo à falta dessa preposição no texto de partida.

Tabela 59 - Testes de múltipla escolha: segmentação e categorização

Cód.	Texto do Teste	Mensagem de treinamento
T19	**A partir da observação do texto de partida, escolha uma só alternativa com a opção mais adequada de tradução em PB:** *We don't accept prepaid financial institutions.* a) Não aceitamos pré-paga instituições financeiras. **b) Não aceitamos instituições financeiras pré-pagas.** c) Não aceitamos instituições financeiras pré-paga. d) Não aceitamos pré-pagas instituições financeiras.	**b) Sugestão de resolução** Em PB, os modificadores geralmente se posicionam após o núcleo do SN e devem concordar com ele em número e gênero.
T25	**A partir da observação do texto de partida, escolha as alternativas com as traduções mais adequadas em PB. Lembre-se que nestes casos é possível selecionar mais de uma opção.** *You have a month to respond to bank inquiries.* a) Você tem um mês para responder ao banco dos inquéritos. **b) Você tem um mês para responder aos inquéritos bancários.** **c) Você tem um mês para responder aos inquéritos do banco.** d) Você tem um mês para responder aos bancários inquéritos.	**b) e c) Sugestões de resolução** Em PB, nomes não podem modificar ou complementar diretamente o núcleo do SN. Por isso, além da mudança na ordem de "*bank*", é necessário inserir uma preposição (c) ou mudar a categoria gramatical desse nome (a).

Tabela 60 - Testes de múltipla escolha: conhecimentos linguísticos

7.2.2 Sugestões de *Golden Text*

Na presente subseção, são sugeridos alguns textos que podem ser utilizados na construção de *Golden Texts,* ou seja, modelos de tradução com erros de tradução previamente controlados, ligados à respostas ideais com propostas consideradas mais adequadas segundo as instruções da empresa e as sugestões tratadas na seção 7.1. É possível avaliar os anotadores e editores a partir da comparação entre o que foi feito por eles durante a avaliação e aquilo que idealmente deveria ter sido feito, presente no *Golden Text*. A vantagem desse tipo de avaliação reside na possibilidade de inserir os textos no verdadeiro contexto da ferramenta com a qual a pessoa avaliada vai trabalhar.

As tabelas apresentadas a seguir possuem a seguinte organização: na primeira coluna, há o texto original na LP e a tradução na LC, contendo os erros controlados; na segunda são apresentadas propostas de anotação[59] (tabela 61) ou de pós-edição (tabela 62) dos erros presentes no texto de chegada; na terceira coluna há uma mensagem de treinamento para o avaliado, caso tenha respondido à avaliação de forma inadequada.

Como no caso do exemplo T25 inserido na tabela 60, o exemplo GT7 inserido na tabela 62 possui erros de tradução que tratam de fenômenos de concordância e ordem de palavras sujeitos à variação em PB ou cujas características fazem com que seja impossível estabelecer uma forma única de traduzir (cf. também Tabela 9.3 do Anexo 9). Por isso foram inseridas duas possibilidades de *Golden Text* nessa tabela, sendo ainda possível criar mais possibilidades de tradução envolvendo os fenômenos tratados.

[59] A anotação da tabela 61 é representada pelos símbolos "[]" e "->" para delimitar a segmentação e categorização, respectivamente.

<table>
<tr><td colspan="4">Golden Text para segmentação e categorização (GT1)</td></tr>
<tr><td colspan="2">Texto de avaliação para Golden Text</td><td>Sugestão de segmentação e categorização</td><td>Mensagem de treinamento</td></tr>
<tr><td colspan="2">A partir da observação do texto de partida, anote os erros presentes na tradução em PB.</td><td rowspan="2">Se você [tens] problemas com o [voos], recomendamos entrar em contato com [diretamente] [passagem] [fornecedor] aérea.

[tens] -> Agreement
[voos] -> Spelling
[diretamente] -> Word Order
[passagem] -> Word Order
[fornecedor] -> Omitted Preposition</td><td rowspan="2">Apesar de “você” ser o controlador da concordância, selecione somente a unidade que contém o erro. No caso de “voos”, essa palavra é singular no texto de partida, por isso somente esse elemento deve ser selecionado e categorizado sob a etiqueta “Spelling”. No caso de erro envolvendo a ordem, selecione somente a menor unidade que, ao ser reposicionada, resolveria o problema. Caso há duas unidades de mesmo tamanho mal poscionadas, selecione a primeira que ocorre no texto.</td></tr>
<tr><td>Texto de partida</td><td>Tradução na língua de chegada</td></tr>
<tr><td>If you have a problem with the flight, we would recommend contacting the airline ticket provider directly. You can also take a look at our Help Center article.</td><td>Se você tens problemas com o voos, recomendam os entrar em contato com diretamente passagem fornecedor aérea.</td><td></td><td></td></tr>
</table>

Tabela 61 – Exemplo GT1: *Golden Text* para segmentação e categorização

<table>
<tr><td colspan="5">Golden Text para avaliar e treinar conhecimentos linguísticos (GT7)</td></tr>
<tr><td colspan="2" rowspan="2">Texto de avaliação para Golden Text</td><td colspan="2">Propostas de pós-edição para Golden Text</td><td rowspan="2">Mensagem de treinamento</td></tr>
<tr><td>Possível tradução 1</td><td>Possível tradução 2</td></tr>
<tr><td colspan="2">A partir da observação do texto de partida, edite os erros presentes na tradução em PB.</td><td rowspan="2">Uma captura de tela anexa do problema pode nos ajudar muito a entender o problema, mas a maioria dos problemas visuais podem ser resolvidos reconectando os seus cabos.</td><td rowspan="2">Uma captura de tela anexa do problema muito pode nos ajudar a entender o problema, mas a maioria dos problemas visuais pode ser resolvida reconectando os seus cabos.</td><td rowspan="2">Os advérbios podem ocupar várias posições em PB. Porém, na ordem apresentada o advérbio “muito” altera o sentido do texto original. Há em PB o fenômeno de concordância semântica. Logo, é possível a concordância gramatical com os traços do núcleo do SN sujeito “maioria” ou semântica com o núcleo do SN mais encaixado “problemas”.</td></tr>
<tr><td>Texto de partida</td><td>Tradução na língua de chegada</td></tr>
<tr><td>An attached screenshot of the issue can help us a lot navigating the problem, but most visual issues can be resolved by reconnecting your cables.</td><td>Uma captura de tela anexa do problema pode nos ajudar a entender muito o problema, mas a maioria dos problemas visuais pode ser resolvido reconectando os seus cabos.</td><td></td><td></td><td></td></tr>
</table>

Tabela 62 – Exemplo GT7: *Golden Text* para conhecimentos linguísticos

7.3 Fóruns de discussão

A partir da discussão dos erros de concordância e ordem de palavras encontrados nos dados (cf. seção 6.2), foi possível verificar a complexidade dos erros, principalmente em casos mais problemáticos, envolvendo, por exemplo, fenômenos de variação. Durante a elaboração das sugestões apresentadas em 7.1 e 7.2, foi possível perceber a dificuldade em elaborar orientações para as *Annotation* e *Language Guidelines* que não sejam demasiadamente longas, mas que tratem de maneira completa os possíveis erros presentes nos textos traduzidos pela empresa.

Considerando-se essas informações, na presente seção se sugere que a empresa implemente em sua plataforma um espaço com "fóruns de discussão" em que os próprios anotadores e pós-editores podem inserir suas dúvidas e ajudar os outros através de suas próprias experiências de tradução na empresa. Esses fóruns também podem ser utilizados pela própria empresa como um canal direto para a discussão com seus pós-editores e anotadores, possibilitando a resolução de dúvidas sobre o processo de anotação ou sobre aspectos linguísticos mais complexos.

8. CONCLUSÃO

Na presente pesquisa, foi possível identificar os erros e confusões mais frequentes na segmentação e categorização dos dados, duas etapas do processo de anotação, envolvendo as etiquetas *Agreement* e *Word Order* no par linguístico inglês-PB. Para melhorar esses processos, foram sugeridas alterações nas *Annotation Guidelines* da empresa que podem promover a uniformidade das unidades categorizados e evitar a anotação de unidades que não contém erro. Quanto às etapas de avaliação e treinamento, foram propostos testes de resposta múltipla e *Golden Tasks,* referidas na presente pesquisa como *Golden Texts*, que podem ser incorporados na plataforma da empresa para avaliar os anotadores e treiná-los através de mensagens fornecidas no final de cada teste.

Também foi possível identificar os casos mais problemáticos e difíceis de controlar envolvendo a ordem de palavras e a concordância em PB. Esses casos problemáticos geralmente envolvem fenômenos como a concordância semântica, a concordância que envolve infinitivos flexionados, a ordem de advérbios e complementos verbais, bem como a posição dos clíticos. Alguns desses fenômenos sintáticos estão sujeitos à variação presente no PB e outros envolvem a subjetividade do pós-editor. Logo, é difícil determinar a naturalidade do excerto traduzido e elaborar regras fixas acerca do funcionamento dos mesmos. Na presente pesquisa, através dos resultados da análise desses fenômenos, foram propostas sugestões para as *Language Guidelines* da empresa que podem promover a reflexão dos pós-editores e anotadores acerca desses fenômenos, bem como salientar a importância deles na tradução de textos. Para a etapa de avaliação, foram sugeridos *Golden Tasks* e testes de múltipla escolha com uma só resposta adequada para casos que possuem regras mais fixas em PB e testes com mais de uma resposta adequada para casos que envolvem os fenômenos mais problemáticos já citados. O treinamento pode ser feito através de mensagens enviadas aos avaliados imediatamente após a escolha da resposta ao teste, fazendo com que o avaliado reflita sobre as escolhas de tradução que acabou de fazer durante o teste.

Finalmente, é sugerida a criação de fóruns de discussão que podem ser utilizados como um canal de comunicação direto entre a empresa e os anotadores e pós-editores. Além disso, através desses fóruns a empresa pode tomar conhecimento acerca das principais dificuldades encontradas e os outros anotadores e pós-editores mais experientes podem auxiliar os outros com as suas próprias habilidades.

Para expandir a pesquisa feita no presente trabalho, é possível aplicar e conferir o impacto das sugestões propostas na qualidade dos textos traduzidos e na uniformidade dos resultados das anotações. Outros aspectos que valem ser examinados com mais profundidade são os fenômenos linguísticos mais variáveis e problemáticos, relacionados com a criatividade do falante e sujeitos à subjetividade do anotador e pós-editor. A análise feita na presente pesquisa também fornece material para a construção de regras para ferramentas de detecção automática de erros, principalmente em casos envolvendo categorias gramaticais e estruturas menos problemáticas relativamente à ordem e a concordância em PB.

9. REFERÊNCIAS

ALPAC. 1966. *Language and Machines: Computers in Translation and Linguistics: A Report*. Washington: National Academy of Sciences, National Research Council. http://site.ebrary.com/id/10071520

Anastasiou, D., e R. Gupta. 2011. "Crowdsourcing as Human-Machine Translation (HMT)". *Journal of Information Science*, 1-25.

Arnold, D., L. Balkan, S. Meijer, R. L. Humphreys, e L. Sadler. 1994. *Machine translation: an introductory guide*. Oxford: Blackwell.

Banerjee, S., e A. Lavie. 2005. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL)*.

Barbosa, P. 2013. "Subordinação argumental finita". In Raposo, E. et al. (coord.). *Gramática do português*, vol. 2, 1821-1897. Lisboa: Fundação Calouste Gulbenkian.

Barbosa, P., e E. Raposo. 2013. "Subordinação Argumental Infinitiva". In Raposo, E. et al. (coord.). *Gramática do português*, vol. 2, 1901-77. Lisboa: Fundação Calouste Gulbenkian.

Bentivogli, L., A. Bisazza, M. Cettolo, e M. Federico. 2016. "Neural versus Phrase-Based Machine Translation Quality: a Case Study". EMNLP 2016. 16 de agosto de 2016. http://arxiv.org/abs/1608.04631

Bechara, E. 1961/2002. *Moderna Gramática Portuguesa*. Rio de Janeiro: Lucena. 37ª edição.

Brito, A. M., e E. Raposo. 2013. "Complementos, modificadores e adjuntos no sintagma nominal". In Raposo, E. et al. (coord.). *Gramática do português*, vol. 1, 1045-1114. Lisboa: Fundação Calouste Gulbenkian.

Cabral, G. de Sousa. 2006. "A Concordância Variável Do Infinitivo Na Escrita Padrão". Mestrado. Universidade Federal do Rio de Janeir o (UFRJ).

Castilho, A. T. de. 2010. *Nova Gramática do Português Brasileiro*. São Paulo: Contexto.

Castilho, S., J. Moorkens, F. Gaspari, I. Calixto, J. Tinsley, e A. Way. 2017. "Is Neural Machine Translation the New State of the Art?". *The Prague Bulletin of Mathematical Linguistics* 108 (1): 109-20.

Castilho, S., J. Moorkens, F. Gaspari, M. Popović, e A. Toral. 2019. “Editors' foreword to the special issue on human factors in neural machine translation”. *Machine Translation* 33 (1-2).

Castro, M. L. de. 2016. “A Variação Na Concordância Verbal: Um Estudo Na Escrita de Acadêmicos de Letras”. Mestrado. Universidade Federal de Rondônia (UNIR).

Comparin, L. 2016. “Quality in machine translation and human post-editing: error annotation and specifications”. Tese de mestrado em tradução, Lisboa: Universidade de Lisboa.

Costa, J. 2008. *O advérbio em português europeu*. Lisboa: Colibri.

Costa-Jussà, M. R., e J. A. R. Fonollosa. 2015. “Latest Trends in Hybrid Machine Translation and Its Applications”. *Computer Speech & Language* 32 (1): 3-10.

Cunha, C., e L. F. L. Cintra. 1985. *Breve gramática do português contemporâneo*. Lisboa: Sá da Costa.

Cunha, F., e M. Moita. 2011. “Advérbios em -mente em estruturas parentéticas”. *Linguagem: Teoria, Análise e Aplicações*, 6: 75-83. Rio de Janeiro: Programa de Pós-graduação em Letras.

Dorr, B. J., P. W. Jordan, e J. W. Benoit. 1999. “A survey of current paradigms in machine translation”. *Advances in computers*, 49: 2.

Duarte, I. 2003. “Relações gramaticais, esquemas relacionais e ordem de palavras”. In Mateus, M. H. et al., *Gramática da Língua Portuguesa*, 275-321. Lisboa: Caminho.

———. 2003. “Padrões de colocação dos pronomes clíticos”. In Mateus, M. H. et al., *Gramática da Língua Portuguesa*, 847-867. Lisboa: Caminho.

Duarte, I., A. L. Santos, e A. Gonçalves. 2016. “O infinitivo flexionado na gramática do adulto e na aquisição de L1”. In Martins, A. M., e E. Carrilho (eds.). *Manual de Linguística Portuguesa*, 453-480. Berlin/Boston: De Gruyter.

———. 2012. “Infinitivo flexionado, independência temporal e controlo”. In Costa A., C. Flores, e N. Alexandre (eds.), *Textos Selecionados do XXVII Encontro Nacional da Associação Portuguesa de Linguística*, 217-244. Lisboa: APL.

Duarte, M., e C. Ribeiro Serra. 2015. “Gramática(s), Ensino de Português e ‘Adequação Linguística’”. *Matraga 22*, 36: 31-55. http://dx.doi.org/10.12957/matraga.2015.17046

Duarte, M. 2015. “Para uma nova descrição da sintaxe do ‘Português Padrão’”. *Cadernos de Letras Da UFF Dossiê: Variação Linguística e Práticas Pedagógicas*, 51: 23-41.

European Commission. 2012. *Crowdsourcing Translation*. Luxembourg: Publication Office of the European Union.

European Commission. *EU General Data Protection Regulation (GDPR)*. https://eugdpr.org/

Faria, A. L. 2017. "Paradoxos no ensino de sintaxe do português". *Linguagem em (Re)vista* 12 (23): 140-165.

Figueira, F. 2018. "Impact of decision tree on annotation system". Tese de bacharelado em tradução, Lisboa: Universidade de Lisboa.

Forcada, M. L. 2017. "Making sense of neural machine translation". *Translation Spaces* 6 (2): 291-309.

Galves, C. 2001. *Ensaios sobre as gramáticas do português*. Campinas: Unicamp.

Geraldo, A. 1998. "Inflected infinitive in romance languages". In *Cad.Est.Ling.* (34): 7-17, Campinas: UniCamp.

Gonçalves, A., e E. Raposo. 2013. "Verbo e sintagma verbal". In Raposo, E., et al. (coord.) *Gramática do português*, 1153-1218. Lisboa: Fundação Calouste Gulbenkian.

Gonçalves, A., A. L. Santos, e I. Duarte. 2014. "(Pseudo-)Inflected infinitives and Control as Agree". In Lahousse, K., e S. Marzo (org.s). *Romance Languages and Linguistic Theory 2012. Selected papers from 'Going Romance' Leuven 2012*: 161-180. Amsterdam/Philadelphia: John Benjamins.

Gornostay, T. 2008. "Machine Translation Evaluation".

Greenbaum, S. 1996. *The Oxford English Grammar.* New York: Oxford University Press.

Groothuis, K. A. 2015. "The inflected infinitive". Romance Research Master thesis in Linguistics.

Han, A. L.-F., D. F. Wong, e L. S. Chao. 2016. "Machine Translation Evaluation: A Survey". https://arxiv.org/pdf/1605.04515v6.pdf

Hutchins, W. J. 2010. "Machine translation: a concise history". *Journal of Translation Studies* 13 (1-2): 29-70.

Hutchins, W. J., e H. L. Somers. 1992. *An introduction to machine translation*. London: Academic.

Johnson, M., M. Schuster, V. Le Quoc, M. Krikun, Y. Wu, Z. Chen, N. Thorat, et al. 2017. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation". *Transactions of the Association for Computational Linguistics*, 5: 339-51.

Kanthack, G. S. 2002. *Clíticos no português brasileiro*. Doutorado. Universidade federal de Santa Catarina, Florianópolis.

Kato, M. A. 2005. *Gramática Do Letrado*. In Marques, M. A. et al. (orgs.). *Ciências Da Linguagem: Trinta Anos de Investigação e Ensino*. Braga: CEHUM.

Kato, M. A., e A. M. Martins. 2016. "European Portuguese and Brazilian Portuguese: An Overview on Word Order". In Wetzels, L., S. Menuzzi e J. Costa (eds.). *The Handbook of Portuguese Linguistics*, 15-40. Hoboken, NJ: Wiley-Blackwell.

Kenny, D. 2018. "Machine translation". In *The Routledge handbook of translation and philosophy*, 428-45. New York: Routledge.

Koehn, P. 2016. "The State of Neural Machine Translation (NMT)". *Omniscien Technologies* (blog). 30 de novembro de 2016. https://omniscien.com/state-neural-machine-translation-nmt/

Koehn, P., e R. Knowles. 2017. "Six Challenges for Neural Machine Translation". In *First Workshop on Neural Machine Translation 2017*, 28-39. Vancouver: Association for Computational Linguistics.

Lobo, M. 2013. "Dependências Referenciais". In Raposo, E. et al. (coord.). *Gramática Do Português*, vol. 2, 2127-2227. Lisboa: Fundação Calouste Gulbenkian.

Lommel, A. 2015. "Multidimensional Quality Metrics (MQM) Definition". QT21 - Quality Translation 21. 30 de dezembro de 2015. http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

Lommel, A., e A. Burchardt. 2014. "Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality". 2014. http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf

Lommel, A., M. Popović, e A. Burchardt. 2012. "Assessing Inter-Annotator Agreement for Translation Error Annotation". In *Automatic and Manual Metrics for Operational Translation Evaluation*, 31-37. http://mte2014.github.io/MTE2014-Workshop-Proceedings.pdf

Lommel, A., H. Uszkoreit, e A. Burchardt. 2014. "Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics". *Tradumàtica: tecnologies de la traducció*, 12: 455-63.

Lucchesi, D. 2012. "A Diferenciação Da Língua Portuguesa No Brasil e o Contato Entre Línguas". *Estudos de Lingüística Galega*, 44: 45-65. https://doi.org/10.3309/1989-578X-

Luís, A. R., e G. A. Kaiser. 2016. "Clitic Pronouns: Phonology, Morphology, and Syntax". In Wetzels, L., S. Menuzzi e J. Costa (eds.). *The Handbook of Portuguese Linguistics*, 210-233. Hoboken, NJ: Wiley-Blackwell.

Luong, T., K. Cho, e C. Manning. 2016. "Neural Machine Translation – Tutorial ACL 2016".

http://nlp.stanford.edu/projects/nmt/Luong-Cho-Manning-NMT-ACL2016-v4.pdf

Martins, A. M., e J. Costa. 2016. "Ordem dos constituintes frásicos: sujeitos invertidos; objetos antepostos". In Martins, A. M., e E. Carrilho (eds.). *Manual de Linguística Portuguesa*, 371-400. Berlin/Boston: De Gruyter.

Martins, A. M. 2019. "OpenKiwi: An Open Source Framework for Quality Estimation". *Medium* (blog). 26 de fevereiro de 2019. https://medium.com/unbabel/openkiwi-an-open-source-framework-for-quality-estimation-30c35a998a9f

———. [s.d.]. "TurboParser (Dependency Parser with Linear Programming)". CMU. Acessado 27 de março de 2019. www.cs.cmu.edu/~ark/TurboParser/

Martins, A. M., M. Almeida, e N. A. Smith. 2013. "Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers". *Conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the Conference* 51: 617-22.

Martins, A. M., M. Junczys-Dowmunt, F. Kepler, R. Astudillo, C. Hokamp, e R. Grundkiewicz. 2017. "Pushing the Limits of Translation Quality Estimation". *Transactions of the Association for Computational Linguistics* 5: 205-18.

Mattos e Silva, R. V. 2013. "O Português do Brasil". In Raposo, E. et al. (coord.). *Gramática do português*, vol. 1, 143-154. Lisboa: Fundação Calouste Gulbenkian.

Mendes, A. 2013. "Organização textual e articulação de orações". In Raposo, E. et al. (coord.). *Gramática do português*, vol. 2, 1691-1819. Lisboa: Fundação Calouste Gulbenkian.

Miguel, M., and E. Raposo. 2013. "Determinantes." In Raposo, E. et al. (coord.). *Gramática do português*, vol. 1, 819-878. Lisboa: Fundação Calouste Gulbenkian.

Monteiro, J. L. 1996. "A Variação Do Infinitivo Em Português". *Revista de Letras* 18 (1): 62-68.

Moorkens, J. 2017. "Under pressure: translation in times of austerity". *Perspectives Studies in Translatology*.

Nowak, S., e S. Rüger. 2010. "How reliable are annotations via crowdsourcing? A study about inter-annotator agreement for multi-label image annotation". In *Proceedings of the international conference on Multimedia information retrieval*, 557-67.

Oliveira, I. de, S. Monguilhott, e I. L. Coelho. 2011. “Entre ordem e concordância”. https://revistas.ufrj.br/index.php/diadorim/article/view/7971

Pagotto, E. G. 1992. “A posição dos clíticos em português: um estudo diacrónico”. Dissertação de Pós-Graduação. Universidade estadual de Campinas, Campinas.

Papineni, K., S. Roukos, T. Ward, e W.-J. Zhu. 2002. “BLEU: A method for automatic evaluation of machine translation”. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–18.

Payne, T. 2010. *Understanding English Grammar*. Cambridge: Cambridge University Press.

Peres, J. A. 2013. “Semântica do Sintagma Nominal”. In Raposo, E. et al. (coord.). *Gramática do português*, vol. 1, 735-818. Lisboa: Fundação Calouste Gulbenkian.

Perini, M. A. 2005. *Gramática Descritiva Do Português*. São Paulo: Ática.

Possenti, S. 1996. *Por Que (Não) Ensinar Gramática Na Escola*. Campinas, SP: Associação de Leitura do Brasil.

“Projeto AC/DC: Corpo NILC/São Carlos.” n.d. https://www.linguateca.pt/acesso/corpus.php?corpus=SAOCARLOS

QT21. 2014. “MQM (Multidimensional Quality Metrics)”. QT21 - Quality Translation 21. 27 de outubro de 2014. http://www.qt21.eu/downloads/MQM-overview-2014-10-27.pdf

———. 2019a. “QT21 - Introduction”. QT21 - Quality Translation 21. 2019. http://www.qt21.eu/?target=Introduction

———. 2019b. “Quality Metrics”. QT21 - Quality Translation 21. 2019. www.qt21.eu/quality-metrics/

QTLaunchpad. 2014a. “Background and Principles”. QT21 - Quality Translation 21. 21 de novembro de 2014. http://www.qt21.eu/launchpad/content/background-and-principles.html

——. 2014b. “High‐Level Structure”. QT21 - Quality Translation 21. 21 de novembro de 2014. http://www.qt21.eu/launchpad/content/high-level-structure-0.html

———. 2014c. “MQM Features”. QT21 - Quality Translation 21. 21 de novembro de 2014. http://www.qt21.eu/launchpad/content/mqm-features.html 1/1

Quah, C. K. 2006. *Translation and technology*. New York: Palgrave MacMilan.

Raposo, E. 1987. “Case Theory and Infl-to-Comp: The Inflected Infinitive in European Portuguese”. *Linguistic Inquiry*, 18 (1): 85-109. MIT Press. http://www.jstor.org/stable/4178525

———. 2013. “Estrutura da Frase”. In Raposo, E. et al. (coord.). *Gramática do português*, vol. 1, 303-428. Lisboa: Fundação Calouste Gulbenkian.

———. 2013. “Advérbio e sintagma adverbial”. In Raposo, E. et al. (coord.). *Gramática do português*, vol. 2, 1573-1684. Lisboa: Fundação Calouste Gulbenkian.

———. 2013. “Pronomes”. In Raposo, E. et al. (coord.). *Gramática do português*, vol. 1, 883-920. Lisboa: Fundação Calouste Gulbenkian.

———. 2013. “Verbos auxiliares”. In Raposo, E. et al. (coord.). *Gramática do português*, vol. 2, 1221-1280. Lisboa: Fundação Calouste Gulbenkian.

Raposo, E., e M. Miguel. 2013. “Introdução ao Sintagma Nominal”. In Raposo, E. et al. (coord.). *Gramática do português*, vol. 1, 703-734. Lisboa: Fundação Calouste Gulbenkian.

Scherre, M. M. P., e A. J. Naro. 1998a. “Restrições Sintáticas e Semânticas No Controle Da Concordância Verbal Em Português”. *Fórum Linguístico, Fpolis* 1: 45-71.

———. 1998b. “Sobre a Concordância de Número No Português Falado Do Brasil”. Ruffino, G. (org.), *Dialettologia, Geolinguistica, Sociolinguistica. (Atti Del XXI Congresso Internazionale Di Linguistica e Filologia Romanza) Centro Di Studi Filologici e Linguistici Siciliani, Universitá Di Palermo.* Tübingen: Max Niemeyer Verlag 5: 509-23.

Silva, C. C., C.-H. Liu, A. Poncelas, e A. Way. 2018. “Extracting In-domain Training Corpora for Neural Machine Translation Using Data Selection Methods”. In *Proceedings of the Third Conference on Machine Translation (WMT)*, 1: 224-31.

Silva, L. R. da. 2017. “As diferenças entre o que se fala e o que se escreve no português do Brasil: a aquisição do clítico se indeterminador e apassivador”. Mestrado. Universidade estadual de Campinas, Campinas.

Testa, I. 2018. “Quality in human post-editing of machine-translated texts: error annotation and linguistic specifications for tackling register errors”. Tese de mestrado em tradução, Lisboa: Universidade de Lisboa.

Turovsky, B. 2016. “Ten years of Google Translate”. *The Keyword* (blog). 28 de abril de 2016. https://www.blog.google/products/translate/ten-years-of-google-translate/

Unbabel. [s.d.]. “Annotation Guidelines - Version 1.2”.

———. 2014. “The Good, the Bad, and the Ugly about Crowd Translation”. *Unbabel Blog* (blog). 11 de fevereiro de 2014. https://unbabel.com/blog/good-bad-ugly-crowd-translation/

———. 2017a. “How do we scale translation quality and speed at Unbabel?” *Unbabel Blog* (blog). 25 de julho de 2017. https://unbabel.com/blog/scale-translation-quality-speed/

———. 2017b. “How do we keep our customers’ data safe at Unbabel?” *Unbabel Blog* (blog). 16 de agosto de 2017. https://unbabel.com/blog/keep-customers-data-safe-unbabel/

———. 2017c. “A closer look at Unbabel’s award-winning translation quality estimation systems”. *Unbabel Blog* (blog). 23 de maio de 2017. https://unbabel.com/blog/unbabel-translation-quality-systems/

———. 2018. “How Unbabel’s ‘Translation as a Service’ will translate everything to human quality”. *Unbabel Blog* (blog). 10 de abril de 2018. https://unbabel.com/blog/translate-human-quality/

———. 2019a. “Language Guidelines – Portuguese (BR)”. Unbabel Support. 2019. https://help.unbabel.com/hc/en-us/articles/115000788474-Language-Guidelines-Portuguese-BR-

———. 2019b. “What is Unbabel?” Unbabel Support. 2019. https://help.unbabel.com/hc/en-us/articles/360003368493-What-is-Unbabel-

Veloso, R., e E. Raposo. 2013. “Adjetivo e Sintagma Adjetival”. In Raposo, E. et al. (coord.). *Gramática do português*, vol. 2, 1359-1496. Lisboa: Fundação Calouste Gulbenkian.

Veloso, R. 2013. “Subordinação Relativa”. In Raposo, E. et al. (coord.). *Gramática do português*, vol. 2, 2059-2134. Lisboa: Fundação Calouste Gulbenkian.

Vicente, G. 2013. “Numerais.”. In Raposo, E. et al. (coord.). *Gramática do português*, vol. 1, 921-948. Lisboa: Fundação Calouste Gulbenkian.

Weaver, W. 1949. “Translation”. In *Machine translation of languages: fourteen essays*. Cambridge, Mass.: Technol. Press of M.I.T.

Wu, Y., M. Schuster, Z. Chen, V. Le Quoc, M. Norouzi, W. Macherey, M. Krikun, et al. 2016. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. arXiv:1609.08144. 26 de setembro de 2016. https://arxiv.org/abs/1609.08144v2

10. ANEXOS

Anexo 1 – Paradigmas de Tradução Automática

Anexo 2 – Erros frequentes nas *Language Guidelines* (PT-BR)

Anexo 3 – Tipologia de erros das *Annotation Guidelines* 1.2

Anexo 4 – Anonimização manual e filtragem de dados

Anexo 5 – Exemplos de erros de categorização em *Agreement* e *Word Order*

Anexo 6 – Processos de filtragem: "Separação" e "Adição" de dados

Anexo 7 – Exemplos de sugestões para a Segmentação

Anexo 8 – Exemplos de testes de múltipla escolha para avaliação e treinamento

Anexo 9 – Exemplos de *Golden Text* para avaliação e treinamento

Anexo 1
Paradigmas de Tradução Automática

1. Paradigmas de Tradução Automática baseados em regras

Segundo Costa-Jussà et al. (2015: 4), os paradigmas baseados em regras (RBMT) usam informações linguísticas, como dicionários monolíngues e bilíngues, combinadas com conhecimento linguístico humano. As regras são desenvolvidas manualmente para transferir o texto na língua de partida para língua de chegada e todo o processo pode ser lido e modificado pelo desenvolvedor. Esse tipo de sistema considera que a tradução é um processo que consiste na análise e representação do significado do texto na LP para possibilitar que o seu equivalente seja gerado na LC (Quah, 2006: 71). Os paradigmas RBMT eram os predominantes antes da virada do milênio, mas a partir desse momento houve o surgimento de paradigmas estatísticos, que passaram a ser referência (Kenny, 2018: 434).

Os sistemas RBMT são geralmente divididos em três tipos: direto, *transfer* e interlíngua. O primeiro também pode ser designado como tradução automática de "primeira geração", tendo em vista ser o tipo de sistema mais antigo. Os sistemas *transfer* e interlíngua se inserem na tradução automática de "segunda geração", tendo surgido após os paradigmas diretos. Devido às características rudimentares dos paradigmas de primeira geração, que faziam uma análise muito superficial da LP, pesquisadores como Quah (2006) não inserem os sistemas diretos dentro do RBMT. Na presente pesquisa, foi escolhido seguir a divisão proposta por Kenny (2018) e Hutchins (2010), sendo assim, os sistemas diretos foram inseridos na seção referente aos RBMT.

1.1 Tradução Direta

Os sistemas diretos eram muito utilizados no período anterior à publicação do relatório feito pela ALPAC. De acordo com Hutchins e Somers (1992: 4), eles eram projetados para um só par linguístico em uma direção determinada, por exemplo, um texto escrito em inglês é traduzido em português, não sendo possível ter o português como língua de partida, pois o sistema é unidirecional. Esse tipo de paradigma não possui etapas intermediárias no processo de tradução. Por isso, o texto é diretamente

processado na língua desejada e somente é analisada a informação estritamente necessária para a produção da tradução (Hutchins e Somers, 1992: 72).

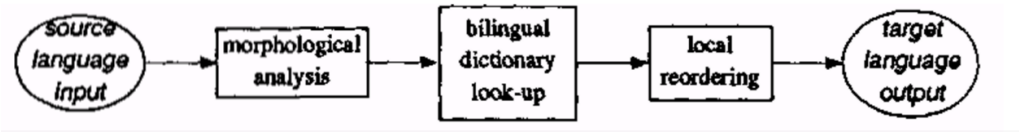

Figura 1.1 – Tradução Direta
(extraída de Hutchins e Somers, 1992: 72)

Como é possível observar na figura 1.1, os sistemas diretos iniciam a tradução com uma fase de análise morfológica do *input*, na qual são identificadas algumas terminações morfológicas das palavras e algumas formas flexionadas são reduzidas para as suas formas não flexionadas. O resultado dessa fase será o *input* de um programa de dicionário bilíngue. Não há a análise da estrutura sintática ou das relações semânticas. O exame feito pelo dicionário bilíngue resulta diretamente em palavras equivalentes. Por isso esse sistema é frequentemente denominado "tradução palavra por palavra". Após essa etapa, o *output* produzido pelo dicionário sofre algumas alterações superficiais de ordenação, produzindo assim o texto de chegada (Hutchins e Somers, 1992: 72).

Segundo Dorr et al. (1999: 13), geralmente a ordem das palavras nesse tipo de tradução é a mesma ordem encontrada no texto da LP. Eles apontam que algumas arquiteturas de sistema direto reconhecem a organização sintática do texto de partida e reorganizam para formas aceitáveis na LC. Todavia, mesmo com essa reorganização superficial, esse tipo de arquitetura produz traduções de baixa qualidade, mas que podem ser úteis para textos simples. Quah (2006: 69) ressalta que esse tipo de método não possui a capacidade de resolver ambiguidades, de lidar com expressões metafóricas ou de traduzir frases entre pares linguísticos distintos.

1.2 Transfer TA

Segundo Kenny (2018: 435), tendo em vista as limitações e os resultados insatisfatórios do método de tradução direta, na segunda geração foram desenvolvidos sistemas que pudessem fazer uma análise inicial da sintaxe e da semântica do texto de partida. Essa análise resultaria numa representação intermediária desse texto que, idealmente, seria mais facilmente traduzido pelo sistema, pois as diferenças entre a LP e a LC teriam sido neutralizadas durante o processo. Dorr et al. (1999: 14) também ressaltam que o objetivo

inicial dos sistemas de arquitetura *transfer* era fornecer textos sintaticamente corretos na língua de chegada, através da transformação de representações sintáticas da LP em representações adequadas na LC.

Nesse tipo de arquitetura a tradução envolve três passos: análise, *transfer* e geração. O primeiro estágio, *análise*, converte o texto numa representação intermediária na qual algumas ambiguidades foram resolvidas. Na segunda etapa, *transfer*, essas representações do texto na língua de partida são convertidas em representações equivalentes da língua de chegada. No terceiro estágio, *geração*, o texto traduzido é produzido.

As etapas de análise e geração são independentes entre elas. A análise feita pelo sistema é dividida por Hutchins e Somers (1992: 4) em análise morfológica, sintática e semântica, sendo a última centralizada na resolução de ambiguidades. A fase de geração também passa por esses três níveis, sendo constituído pela geração morfológica, sintática e semântica. As diferenças de vocabulário e estrutura entre a LC e a LP são tratadas na fase intermediária do processo (Hutchins e Somers, 1992: 4).

À vista disso, a função do módulo de *transfer* bilíngue é converter as representações da LP em representações na LC. Ainda segundo esses autores, "in the transfer approach there are therefore no language-independent representations: the source language intermediate representation is specific to a particular language, as is the target language intermediate representation (Hutchins e Somers, 1992: 75). A figura 1.2 a seguir demonstra essas três etapas do método *transfer* para a tradução de um texto entre francês e inglês:

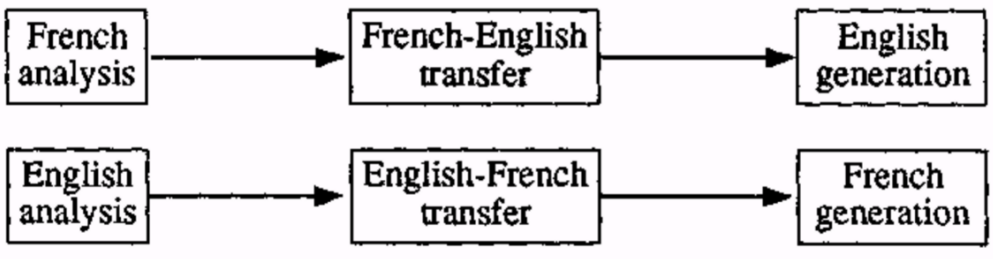

Figura 1.2 – Método *Transfer*
(extraída de Hutchins e Somers, 1992: 75)

Como é possível observar na figura 1.2, o "French-English *transfer*" é específico para a tradução desse par linguístico em que o francês é a LP e o inglês é a LC. Para que a

tradução seja feita na outra direção (inglês como LP e francês como LC), seria necessária a utilização de outro *transfer*, representado na figura como "English-French *transfer*". Em relação a essa característica, Quah ressalta que "unlike the interlingua approach where only one interlingua is responsible for all the language pairs, the transfer approach uses different transfer models for each language pair" (2006: 74).

1.3 Interlíngua

A ideia básica da arquitetura de interlíngua é que a análise do texto na língua de partida resulte numa representação do texto que seja independente dessa língua (Dorr et al., 1999). Esse método busca uma representação abstrata independente e universal, refletindo os objetivos de linguistas teóricos dos anos 1960, como aponta Quah (2006: 71). Idealmente, a interlíngua seria então o estágio intermediário entre todas línguas naturais. Esse tipo de paradigma envolveria uma notação universal, capaz de expressar o significado da língua de partida em qualquer língua de chegada abrangida pelo sistema. Consequentemente, a tradução do texto analisado poderia ser gerada em qualquer língua de chegada.

Apesar dos esforços de inúmeros grupos de pesquisadores em desenvolver a interlíngua nas décadas de 1970 e 1980, Kenny (2018) sustenta que esse paradigma nunca foi utilizado em larga escala e investigadores consideram que ela é impraticável e teoricamente inadequada.

O sistema interlíngua assume a possibilidade de assimilar e produzir o texto a partir de representações comuns a mais de uma língua. A tradução é feita em dois estágios: a partir da LP para a interlíngua e a partir da interlíngua para a LC. Os programas que fazem a análise são independentes dos programas de geração, por isso numa configuração multilíngue qualquer programa de análise pode ser conectado a qualquer outro programa de geração (Hutchins e Somers, 1992: 4). Isso pode ser observado na figura seguinte em que a mesma interlíngua pode ter o inglês ou o francês como língua de chegada (ou de partida). Logo, diferente do *transfer*, esse tipo de arquitetura pode traduzir em ambas as direções.

ANEXO 1 – PARADIGMAS DE TRADUÇÃO AUTOMÁTICA

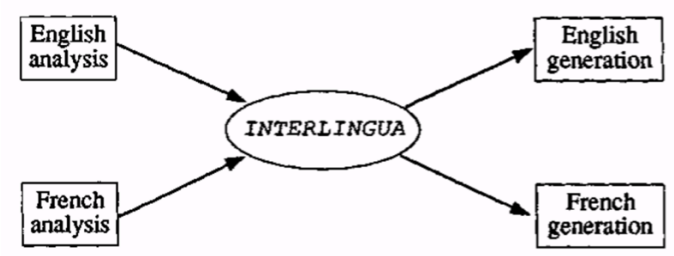

Figura 1.3 – Método Interlíngua

(extraída de Hutchins e Somers, 1992: 74)

Enquanto em *transfer* as regras variam amplamente de língua para língua, sendo então utilizadas para um par linguístico específico, em interlíngua, a mesma representação interna é utilizada para ambas a LC e LP (Dorr et al., 1999: 14). Idealmente, na interlíngua as representações da LC e da LP são idênticas e a gramática comparativa entre as duas línguas não é necessária (Arnold et al., 1994: 81). Nesse caso, a representação produzida pela análise poderia ser diretamente o *input* do componente que faz a síntese da LC, pois essa representação inclui toda a informação necessária para a geração do texto de chegada, sem a necessidade voltar para o texto original. Essa representação é ao mesmo tempo a projeção do texto de partida e a base para a formação do texto na LC. Hutchins e Somers (1992: 73) a definem como sendo "an abstract representation of the target text as well as a representation of the source text".

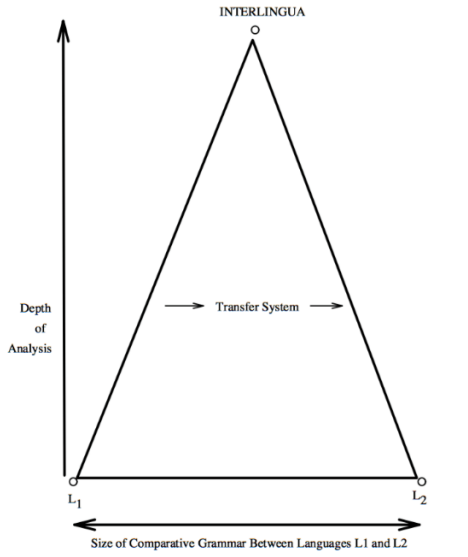

Figura 1.4 – *Transfer* e Interlíngua

(extraída de Arnold, 1994: 81)

Para finalizar, a figura 1.4 ilustra a relação entre a profundidade da análise feita pelo sistema e as arquiteturas *transfer* e interlíngua. Segundo Arnold et al. (1994: 81), a quantidade de gramática comparativa necessária para traduzir entre duas línguas diminui à medida em que a "profundidade" da representação aumenta. À medida que a representação se torna mais abstrata, há menos diferenças entre as representações de LP e LC e fica mais fácil relacioná-las. A interlíngua, representada pelo topo da pirâmide, necessita de uma grande análise da estrutura do texto, mas o uso da gramática comparativa entre LP (L1) e LC (L2) é, idealmente, mínimo, pois essa representação é altamente abstrata. Já no *transfer*, representado pela parte intermediária da pirâmide, a profundidade da análise é variável e a quantidade de gramática comparativa utilizada também varia.

2. Paradigmas de Tradução Automática baseados em corpora

O princípio básico dos *corpus-based machine translation* (CBMT) é que o conhecimento necessário para traduzir pode ser apreendido a partir de *corpora* paralela (ou bitextos), ou seja, a coleção de textos de partida alinhados com as suas respectivas traduções humanas (Kenny, 2018: 435). Esses paradigmas orientados para os dados usam informações retiradas dos dados e complexos algoritmos que juntos são capazes de modelar uma tradução (Costa-Jussà e Fonollosa, 2015: 4). Segundo Dorr et al. (1999: 30), as investigações sobre esse tipo de paradigma foram possíveis devido aos rápidos avanços em computação, além da disponibilidade de dicionários eletrônicos e de *corpora* monolíngue e bilíngue. Esses paradigmas dependem da existência de abundante *corpora* textual, usados para o treinamento da máquina. Esse tipo de paradigma pode ser dividido em três sub-tipos principais: *example based machine translation* (EBMT), *statistical machina translation* (SMT) e *neural machine translation* (NMT). Esse último tipo de sistema não será tratado no presente anexo, tendo em vista já ter sido descrito na seção 2.2 da pesquisa.

2.1 EBMT: Sistemas de Tradução Automática baseados em exemplos

O sistema EBMT foi proposto pela primeira vez em meados dos anos 1980 por Makoto Nagao e sua ideia básica é a utilização de um banco de dados de traduções paralelas, do

qual são extraídas frases na LP já traduzidas anteriormente, que correspondem o máximo possível com as novas frases que ainda devem ser traduzidas (Dorr et al., 1999: 33). Esse tipo de tradução é designada por esses pesquisadores como "Tradução por Analogia", mas ela também é caracterizada como "analogy-, memory-, pattern-, case- or similarity-based translation" (Quah, 2006: 81). Esse sistema era muito semelhante às memórias de tradução, utilizadas em *computer-aided translation*, e foram rapidamente suplantadas por sistemas SMT (Kenny, 2018: 435).

Esse tipo de sistema envolve a identificação e a extração automática a partir dos bitextos de blocos equivalentes no texto de partida e chegada. Esses blocos equivalentes, denominados "example", são então armazenados em bancos de dados (Quah, 2006: 81). Para traduzir uma sentença, o sistema procura uma tradução equivalente do bloco do texto de partida no *corpus* previamente armazenado, através de algoritmos. A tradução das frases escritas na LP que se correspondem é modificada e combinada para formar a tradução da nova frase.

Esse processo pode ser resumido em três etapas: a comparação de segmentos do novo texto escrito na LP com pares de exemplos extraídos previamente do *corpus* bilíngue, depois o alinhamento de segmentos correspondentes e a recombinação desses segmentos para gerar o texto de chegada (Quah, 2006: 81).

2.2 SMT: Sistemas de Tradução Automática baseados em estatística

Os sistemas de tradução automática estatísticos (SMT) produzem traduções a partir da previsão estatística e dependem fortemente da análise estatística de *corpora* paralela bilíngue (Dorr et al., 1999: 30). O primeiro esforço moderno para a construção de um sistema SMT foi iniciado pelo IBM em 1988 com seu Candide *French-English Machine Translation Project* e esse tipo de sistema continua a ser utilizado nos dias atuais.

Segundo Costa-Jussà e Fonollosa (2015: 4), a principal diferença entre os paradigmas EBMT e os SMT é que, por definição, o primeiro faz uma tradução direta por analogia e pode ser considerado com um "*pattern matching problem*", enquanto o segundo tenta encontrar a tradução mais provável a partir dos modelos construídos através de dados.

Em SMT, as informações recolhidas em *corpora* monolíngues e bilíngues ficam separadas, pois a informação monolíngue está localizada no modelo da língua alvo enquanto a informação bilíngue provém do modelo de tradução (Quah, 2006: 80). Esse tipo de sistema se baseia em dois modelos probabilísticos: um é assimilado a partir de

corpora paralela e outro é feito a partir da LC, tendo como base uma extensa quantidade de textos monolíngues (Kenny, 2018: 435). O algoritmo seleciona a tradução considerada como a mais provável, tendo em vista a combinação das informações fornecidas por esses dois modelos (Quah, 2006: 78).

A arquitetura dos sistemas SMT pode ser dividida em três partes: alinhamento, cálculo das correspondências e reordenação (Kenny, 2018: 436). Na primeira fase (*learning*), o sistema aprende os modelos. A seguir, na fase de cálculo das correspondências (*tuning*), os desenvolvedores do sistema investigam qual é o peso ideal que deve ser atribuído para cada modelo a fim de ter o melhor resultado. Na última fase, reordenação (*decoding*), quando o sistema é usado para traduzir uma sentença, ele gera inúmeras traduções hipotéticas da frase na LC e calcula qual delas é a mais provável a partir dos modelos aprendidos na primeira fase e dos pesos atribuídos a eles. Essa fase de atribuição de pesos é essencial nos sistemas SMT e envolve muita investigação para encontrar o peso ideal para os diferentes modelos.

Ainda segundo essa autora, os modelos originais de tradução em SMT propostos pela IBM eram baseados em "*unigrams*", ou seja, uma única palavra. Posteriormente foram desenvolvidas novas técnicas que permitiam o aprendizado de traduções de *n*-grams, ou seja, de cadeias de uma, duas, três ou *n* palavras que apareciam sucessivamente no texto. Essa autora aponta que o número limitado de textos usados para construir os modelos e o fato de que os *n*-grams são traduzidos independentemente um dos outros, não necessariamente correspondendo a nenhum tipo de unidade estrutural na teoria linguística, fazem com que os sistemas SMT tenham grande dificuldade em traduzir termos com relações de dependência discontínua, como a relação entre "*send*" e "*back*" em "*Send your certificate of motor insurance back*" (Kenny, 2018: 436). Esses sistemas também possuem dificuldades em lidar com línguas que possuem um variado sistema de flexão ou aglutinação, pois não possuem orientações sobre como lidar com a concordância gramatical (Kenny, 2018: 436).

3. Paradigmas de Tradução Automática híbridos

É de notar a existência dos chamados sistemas "híbridos", que combinam as melhores características de mais de um paradigma. Costa-Jussà e Fonollosa (2015: 5) ressaltam a existência de duas possibilidades: a primeira são os sistemas híbridos guiados por RBMT que integram informações obtidas a partir dos dados em arquiteturas *rule-based*, a

segunda são os sistemas híbridos guiados por CBMT que integram regras linguísticas em sua arquitetura *corpus-based*.

Quah (2006: 84-85) ressalta as características desses dois tipos de paradigmas que levam ao uso de sistemas híbridos. Para ele, apesar dos sistemas *rule-based* serem dedutivos e baseados em regras linguísticas estabelecidas pelos seus desenvolvedores, eles não guardam seus resultados de tradução ou reutilizam os segmentos traduzidos anteriormente. Já os sistemas *corpus-based* são mais indutivos, pois suas regras provêm de um grupo pré-estabelecido de exemplos de tradução e a modificação dessas regras é feita a partir da inserção de novos exemplos de tradução. Tendo em vista essas e outras características, Quah considera que é pouco provável que a pesquisa em tradução automática continue a progredir se um paradigma for preferido em relação ao outro (Quah, 2006: 85). Logo, os sistemas híbridos e outros paradigmas inovativos seriam a melhor forma de avançar na tradução automática.

Anexo 2
Erros frequentes nas *Language Guidelines* (PT-BR)

Erros mais frequentes – *Language Guidelines*						
Nome da seção	**Nome das subseções**					
Erros de exatidão	Tradução literal			Seleção lexical		
Erros de fluência	Ortografia	Uso de letras maiúsculas e minúsculas	Pontuação	Artigos desnecessários / ausentes	Tempos verbais	Ordem das palavras
Erros de estilo	Registro					
Terminologia	Não conformidade com as instruções do cliente ou da empresa			Não conformidade com o glossário e com o vocabulário		
Variação incorreta na mesma língua						
Nomes de entidades	Moeda	Data / hora	Siglas e acrônimos		Traduções Consagradas	

Tabela 2.1 – Erros mais frequentes nas *Language Guidelines*

(informações extraídas de Unbabel, 2019a)

ANEXO 2 – ERROS FREQUENTES NAS *LANGUAGE GUIDELINES* (PT-BR)

Nome da subseção	**Sugestão**	**Exemplos em inglês (EN), português brasileiro (PT ou PT-BR) e português europeu (PT-EU)**
Tradução literal	Atenção para evitar a tradução literal de expressões idiomáticas, ou mesmo de palavras.	EN: *Sorry to hear that.* PT (literal): *Lamento ouvir isso.* PT (correto): *Sinto muito por isso.*
Ordem das palavras	A ordem das palavras não é a mesma que se utilizaria na língua de chegada	EN: *It is illegal to **artificially change** the number of visits to the site.* PT (incorreto): *É ilegal **artificialmente alterar** o número de visitas ao site.* PT (correto): *É ilegal **alterar artificialmente** o número de visitas ao site.*
Variação incorreta da mesma língua	Você deve ter em conta a variação linguística entre o português do Brasil e o português europeu. A variação lexical (utilização de palavras e expressões diferentes e cujo uso mais comum pode variar) e morfossintática (um dos maiores problemas está relacionado com a ordem das palavras) deve ser detectada e contornada sempre. Por favor, busque soluções adequadas ao português do Brasil. No que diz respeito a problemas morfossintáticos, é preciso estar atento a diferenças verbais (ordem das palavras e concordância).	EN: *He gave me a book.* PT-EU: ***Ele ofereceu-me** um livro.* PT-BR: ***Ele me ofereceu (ou ofereceu-me)** um livro.*

Tabela 2.2 – Subseções de erros frequentes
(informações extraídas do site da Unbabel, 2019a)

Anexo 3
Tipologia de erros das *Annotation Guidelines* 1.2

ACCURACY: O texto de chegada não é fiel ao texto de partida.		
Addition:	O texto de chegada inclui uma unidade não presente no original.	
Omission: O conteúdo presente no original está em falta no texto de chegada ou o conteúdo está em falta somente no texto de chegada.	*Omitted Preposition*	Omissão de uma preposição no texto de chegada.
	Omitted Conjunction	Omissão de uma conjunção no texto de chegada.
	Omitted Determiner	Omissão de um determinante no texto de chegada.
	Omitted Pronoun	Omissão de um pronome no texto de chegada.
	Omitted Auxiliary Verb	Omissão de um verbo auxiliar no texto de chegada.
	Other POS Omitted	Omissão de uma ou mais palavras pertencentes a qualquer categoria morfológica (com exceção das mencionadas acima) no texto de chegada.
Mistrans-lation: O conteúdo no texto de chegada não representa corretamente o texto de partida.	*Ambiguous Translation*	Um texto não ambíguo foi traduzido de maneira ambígua no texto de chegada.
	Named Entity	Nomes, lugares, localidades ou outras entidades mencionadas não se correpondem entre o texto de partida e chegada.
	False Friend	A tradução utilizou incorretamente uma palavra que é graficamente similar a outra na língua de partida, mas que possui um significado completamente diferente na língua de chegada.
	Overly Literal	A tradução está excessivamente restrita ao texto de partida, o que pode causar problemas de interpretação.
	Lexical Selection	O termo selecionado não está correto para o contexto ou é inexato para transmitir o sentido do texto original.
	Shouldn't have been translated	Foi traduzido um texto que não deveria ter sido traduzido.
	Spelling	Erros de ortografia que envolvem o uso de palavras similares na língua de chegada que possuem significados distintos e também quando uma forma no plural é utilizada ao invés da forma no singular ou vice-versa.
	Date/Time	Datas ou horários não correspondem entre o texto de partida e de chegada devido a diferentes formatos.
	Number	Números são inconsistentes entre o texto de chegada e de partida.

	Unite Convertion	Erros que envolvem a substituição de unidades de conversão.
Over-translation	O texto de chegada é mais específico do que o texto de partida.	
Under-translation	O texto de chegada é menos específico do que o texto de partida.	
Untranslated	Não foi traduzido um conteúdo que deveria ter sido traduzido.	

Tabela 3.1 – Tipologia de erros de *Accuracy*

(informações baseadas nas *Annotation Guidelines* 1.2)

FLUENCY: Problemas que afetam a leitura e a compreensão do texto.			
Character Encoding	Caracteres estão distorcidos devido a aplicação incorreta de uma codificação.		
Duplication	Uma palavra ou uma parcela do texto foi acidentalmente repetida.		
Grammar: Problemas relacionados com a gramática ou a sintaxe do texto sem afetar o significado original do texto.	*Function Word:* Uma categoria gramatical é incorretamente utilizada.	Wrong Preposition	Uma preposição é utilizada incorretamente.
		Wrong Conjunction	Uma conjunção é utilizada incorretamente.
		Wrong Determiner	Um determinante é utilizado incorretamente.
		Wrong Pronoun	Um pronome é utilizado incorretamente.
		Wrong Auxiliary Verb	O verbo auxiliar está incorreto.
	Wordform: Há um problema na estrutura da palavra	Agreement	Uma ou mais palavras não concordam relativamente ao caso, número, pessoa ou gênero.
		Tense/ Mood/ Aspect	Uma forma verbal apresenta o incorreto tempo, modo ou aspecto.
		POS	Uma palavra têm a categorial gramatical incorreta.
	Word Order	A ordem das palavras está incorreta na língua de chegada.	
Inconsistency	O texto de chegada possui inconsistências internas relativamente à tradução dos mesmos termos ou ao uso de abreviações.		
Typography: Problemas relacionados com a apresentação do texto.	*Capitalization*	Uso incorreto de letras maiúsculas ou ausência de letras maiúsculas.	
	Diacritics	Problemas relacionados com o uso de diacríticos.	
	Hyphenation	Uso incorreto de hífen (o hífen no texto de chegada foi usado incorretamente, está em falta ou há um hífen a mais).	

	Othography	Palavras grafadas incorretamente. Essa categoria não se aplica a erros de diacríticos.
	Punctuation	A pontuação ou um dos elementos do par de aspas, parênteses ou pontuação foram usadas incorretamente ou está em falta.
	Whitespace	Um espaço em branco foi utilizado incorretamente (ele está em falta ou há um em excesso.
Unintelligible	A natureza exata do erro não pode ser determinada, demonstrando uma grave quebra na fluência do texto.	

Tabela 3.2 – Tipologia de erros de *Fluency*

(informações baseadas nas *Annotation Guidelines* 1.2)

STYLE: O texto possui erros estilísticos.	
Grammatical Register	O texto usa formas pronominais e verbais que não são compatíveis com o registro solicitado pelo cliente para quele texto específico.
Lexical Register	O texto usa expressões lexicais que não são compatíveis com o registro solicitado pelo cliente para aquele texto específico.
Wrong Language Variety	A variedade linguística utilizada não é a correta.
Company Terminology	O texto não é compatível com as orientações terminológicas do cliente, especificadas no glossário.
Unidiomatic	O conteúdo é gramatical, mas não é idiomático.

Tabela 3.3 – Tipologia de erros de *Style*

(informações baseadas nas *Annotation Guidelines* 1.2)

Anexo 4
Anonimização manual e filtragem de dados

Devido a questões de confidencialidade e respeito às regras do Regime Geral de Protecção de Dados, antes de iniciar o processo de análise e contagem de dados, foi necessário fazer a anonimização manual das informações consideradas sensíveis, geralmente correspondendo a entidades mencionadas. A anonimização feita nesta pesquisa difere da etapa de anonimização automática já referida na seção 3.1 da pesquisa, mas é possível encontrar nos dados recebidos termos que passaram pela anonimização automática como os termos "EMAIL-0" e "URL-0", que correspondem a um endereço de *e-mail* e a um endereço de *website*, respectivamente.

Tendo em vista o objetivo da presente pesquisa, a anonimização dos dados de anotação fornecidos pela empresa foi feita manualmente, preservando-se as informações morfológicas ou lexicais consideradas necessárias para a análise dos erros. Para exemplificar, apresentamos a seguir o caso de uma palavra anonimizada manualmente dentro do seu contexto no texto de partida e no de chegada:

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
A8	*a Bank Statement from a recognised bank in the* ***(COUNTRY M)*** *(not e-money/pre-paid financial institutions);*	*um Extrato Bancário de um banco reconhecido no* ***(COUNTRY M)*** *(não e-money / [pré]-pago [instituições] [financeiras]);*	instituições/financeiras/pré

Tabela 4.1 – Exemplo de dado anonimizado manualmente

O termo assinalado no trecho A8[1], "(COUNTRY M)", corresponde à palavra anonimizada no texto de partida e no texto de chegada. Nesse caso, a palavra original correspondia lexicalmente ao nome de um país (representado pelo termo COUNTRY) e possuía o gênero masculino (representado pela letra M). Nos casos em que foi necessário anonimizar uma grande quantidade de informação, a expressão (CONFIDENTIAL

[1] Os códigos dos segmentos apresentados neste anexo se referem ao número do dado na listagem original, disponibilizada ao júri avaliador: a letra "A" corresponde a exemplos categorizados como *Agreement* pelos anotadores da Unbabel e a letra "W" corresponde a trechos categorizados em *Word Order* pelos mesmos anotadores.

INFORMATION) foi inserida. As informações relativas ao gênero e ao número da palavra somente foram inseridas quando consideradas necessárias para a análise. Nesse caso, o gênero feminino foi representado pela letra “F”, o gênero masculino pela letra “M”, o plural foi representado pelas letras “PL” e o singular foi representado através da ausência de letras após o nome anonimizado correspondente à informação lexical.

É apresentada uma lista dos termos anonimizados manualmente na tabela 4.2 a seguir com informações gerais que podem auxiliar na compreensão do significado lexical correspondente aos termos. É possível observar que a maior parte desses termos corresponde a nomes próprios, sendo raros os casos em que foi necessário anonimizar outra categoria gramatical.

ANEXO 4 – ANONIMIZAÇÃO MANUAL E FILTRAGEM DE DADOS

Termo após anonimização	Exemplo no contexto de partida	Exemplo no contexto de chegada	Categoria gramatical	Informação lexical
(PERSON'S NAME)	*Dear (PERSON'S NAME F),*	*Querido (PERSON'S NAME F),*	Nome próprio	Nome próprio de um indivíduo.
(COUNTRY)	*You (COUNTRY M) or EUR Bank Account number*	*O seu número da contabancária no (COUNTRY M) ou em EUR*	Nome próprio	Nome de um país.
(NATIONALITY)	*...a collaboration between three main (NATIONALITY) TV networks...*	*...uma colaboração entre três principais redes TV (NATIONALITY F PL)...*	Adjetivo	Nacionalidade.
(COMPANY)	*A (COMPANY M) Account is for personal use only...*	*As Conta de (COMPANY M) são apenas para uso pessoal...*	Nome próprio	Nome de uma empresa.
(PRODUCT)	*...select the Cancel (PRODUCT) option in the middle of the page.*	*...selecione a opção Anular (PRODUCT) no meio da página.*	Nome próprio	Nome de um produto da empresa.
(BRAND)	*...we've introduced in partnership with (BRAND M) TV...*	*...introduzimos em parceria com o (BRAND M) TV...*	Nome próprio	Nome de uma marca.
(CARD BRAND)	*that accepts debit (CARD BRAND).*	*que aceite o débito (CARD BRAND).*	Nome próprio	Nome de um marca do cartão de crédito ou débito.
(DOC FORMAT)	*...you can submit the original (DOC FORMAT)...*	*..você pode enviar o formato original (DOC FORMAT)...*	Sigla de nome comum ou de expressão nominal mais complexa	Sigla de um nome relativo ao formato de um documento digital como .doc, .pdf, .xls, .txt., etc.
(NUMBER)	*...the monthly charge for $(NUMBER).*	*...a cobrança mensal de $(NUMBER).*	Numeral	Número, geralmente correspondendo ao preço de um produto no caso dos dados.

Tabela 4.2 – Lista dos termos anonimizados manualmente

ANEXO 4 – ANONIMIZAÇÃO MANUAL E FILTRAGEM DE DADOS

Em seguida, foi feita uma filtragem de dados, através da qual foram retirados os dados com IIA e os dados repetidos. A não-inclusão dos dados IIA não significa que foram mal anotados, mas que, devido à anonimização ou à falta de mais contexto, foi impossível continuar a análise. A tabela 4.3, apresentada a seguir, procura exemplificar alguns desses casos. Em A168, o texto de chegada foi demasiadamente alterado, sendo impossível compreendê-lo. Já em A90, o termo "*obrigado*" foi o segmento categorizado sob a etiqueta *Agreement*, mas não há nenhuma informação nos trechos recebidos que indique o gênero do falante que usa a expressão, sendo impossível verificar a concordância no segmento (cf. seção 5.1.3 da pesquisa, sobre concordância e coesão textual). No caso dos trechos em W61, com segmentos classificados como contendo erro *Word Order,* foram anonimizados dois termos inseridos no mesmo SN, dificultando assim a análise do erro. O caso de A99 é particular, pois foi o único caso encontrado nos dados em que o texto de partida foi escrito em PB, também impossibilitando a análise, pois não envolvia tradução.

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
A168	*(PERSON'S NAME F)*	*[X[x[xxx]]]*	(PERSON'S NAME F)/Obrigado/Obrigado
A90	*Many thanks,*	*Muito [obrigado],*	obrigado
W61	*On the (PRODUCT F) (COMPANY M) app, go to Settings >> Login.*	*No [(PRODUCT F)] [aplicativo] [(COMPANY M)], vá para Configurações > > Entrar.*	aplicativo/(COMPANY M)/(PRODUCT F)
A99	*...por esse motivo solicitamos a ativação e validação desta conta.""*	*...por este motivo [solicita] uma ativação e validação desta conta"."*	solicita

Tabela 4.3 – Exemplos de dados IIA

Os dados repetidos são aqueles que contêm exatamente a mesma informação nos textos de partida e de chegada, bem como na severidade e nos segmentos anotados. Foram encontrados dados repetidos de uma a cinco vezes, totalizando 34 dados repetidos em *Agreement* (10 dados repetidos 10 vezes; 2 dados repetidos 4 vezes; e 1 dado repetido 6 vezes) e 25 dados repetidos em *Word Order* (6 dados repetidos 6 vezes; 1 dado repetido 3 vezes; e 2 dados repetidos 5 vezes), tendo sido mantido nas contagens somente um exemplar de cada um desses dados, logo 13 dados em *Agreement* e 9 dados em *Word*

Order. Para exemplificar, é possível observar na tabela 4.4 que ambos os dados possuem o mesmo texto na LP e LC, tendo sido anotados sob a etiqueta *Word Order* e com a severidade *major*. Além disso, os segmentos anotados também são idênticos em ambas as linhas: as palavras "botão/verde/Baixar" foram selecionadas pelos anotadores. Nesse caso, somente a primeira linha (W28) foi mantida como dado válido para análise.

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Categoria selecionada pelos anotadores da Unbabel**	**Severidade selecionada pelos anotadores da Unbabel**
W28	*Choose the file type that you need, then click the green Download button.*	*Escolha o tipo de arquivo que você precisa e, em seguida, clique no botão [Baixar] verde.*	botão/verde/ Baixar	Word Order	major
-	*Choose the file type that you need, then click the green Download button..*	*Escolha o tipo de arquivo que você precisa e, em seguida, clique no botão [Baixar] verde.*	botão/verde/ Baixar	Word Order	major

Tabela 4.4 – Exemplos de dados repetidos

Há ainda casos em que a categoria anotada e os textos de chegada e de partida são os mesmos, mas uma ou mais informações são diferentes. Na tabela 4.5, apesar de os textos serem os mesmos e de terem sido anotados sob a mesma categoria, foram selecionados segmentos diferentes para cada um dos dados: "ícone/Configurações" em W80 e "no/ícone/Configurações" em W82. O grau de severidade anotado também é diferente: *minor* em W80 e *major* em W81. Em casos destes, os dados não foram considerados como repetidos, pois as divergências entre as informações acerca da segmentação e da severidade serão observadas na subseção 6.2, acerca dos erros relativos ao processo de anotação.

ANEXO 4 – ANONIMIZAÇÃO MANUAL E FILTRAGEM DE DADOS

Código do segmento	Texto de partida	Texto de chegada	Segmentos selecionados pelos anotadores da Unbabel	Categoria selecionada pelos anotadores da Unbabel	Severidade selecionada pelos anotadores da Unbabel
W80	*Tap on the Settings icon at the top right corner*	*Toque no Configurações ícone no canto superior direito*	ícone/Configura-ções	Word Order	minor
W81	*Tap on the Settings icon at the top right corner*	*Toque no Configurações ícone no canto superior direito*	no/ícone/Configu-rações	Word Order	major

Tabela 4.5 – Exemplos de dados semelhantes, mas não repetidos

Anexo 5
Exemplos de erros de categorização em *Agreement* e *Word Order*

Categorias sugeridas na presente análise	**Nº de dados inseridos nestas categorias**	**Exemplo**			
		Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
Agreement / Diacritics	1	A136	*The last 4 digits of the card...*	*[o] últimos 4 digitos do cartão...*	o
Agreement / Word Order	1	A141	*...including secret boards because...*	*... incluindo [secreto] painéis...*	secreto
Agreement / Wrong Determiner	1	A6	*A (COMPANY M) Account is...*	*[As] Conta de (COMPANY M) são...*	As
Agreement /Over-Translation	1	A57	*If you can provide further details about the flights...*	*Se você puder fornecer mais detalhes sobre [o] [passagens] aéreas...*	o / passagens

Tabela 5.1 – Exemplos de categorização de *Agreement* com uma outra etiqueta

ANEXO 5 – EXEMPLOS DE ERROS DE CATEGORIZAÇÃO EM *AGREEMENT* E *WORD ORDER*

Categorias	**Nº de dados inseridos nestas categorias**	**Exemplos**				
		Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Propostas de tradução**
Agreement / Orthography / Punctuation	1	A3	*(5) Screenshot(s) of the issue are very helpful!*	*(5) [Imagem] De Captura De Tela(s) da questão são muito úteis!*	Imagem	(5) Imagem(ns) de captura de tela da questão...
Agreement / Overly Literal / Orthography / Punctuation OU Overly Literal / Punctuation	1	A101	*On your next download, purchase/s will be deducted...*	*No seu próximo baixar, [a compra / s] será deduzida...*	a compra / s	... a(s) compra(s) será(ão) deduzida(s)... OU ...a compra será deduzida...
Agreement / Overly Literal / Orthography / Punctuation OU Não há erro	1	A103	*On your next download, purchase/s will be deducted...*	*No seu próximo baixar, a compra / s [será] deduzida...*	será	... a(s) compra(s) será(ão) deduzida(s)... OU ...a compra será deduzida...

Tabela 5.2 – Exemplos de categorização de *Agreement* com mais de uma etiqueta

Categorias	**Nº de dados inseridos nestas categorias**	**Exemplo**			
		Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
Addition	1	A64	*...take a look at our (PRODUCT M) help center article...*	*... olhe a nossa Artigo [da] ajuda da engrenagem de (PRODUCT M)...*	da
Inconsistency	1	A144	*...you should be the only person who knows your login details and that you are the only one connected...*	*...deve ser a única pessoa que conhece os seus dados de login e que é [o único] ligado...*	o único
Lexical Selection	3	A145	*...and that you are the only one connected to your account.*	*...e que é o único [ligado] à sua conta.*	ligado

Tabela 5.3 – Exemplos de categorização em que há outro tipo de erro (*Agreement*)

ANEXO 5 – EXEMPLOS DE ERROS DE CATEGORIZAÇÃO EM *AGREEMENT* E *WORD ORDER*

Categoria	**Nº de dados inseridos nestas categorias**	**Exemplo**				
		Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Proposta de tradução**
Word Order / Omitted Preposition	1	W76	*Please send them a billing ticket...*	*Por favor enviar eles um [cobrança] [ingresso]...*	cobrança/ ingresso	...um ingresso de cobrança...
Word Order / Agreement	2	W15	*Add a secondary email address (URL-0)...*	*Adicione um endereço de [e-mail] ail (URL-0 tings) de [secundária]*	e-mail/ secundária	Adicione um endereço de e-mail secundário (URL-0)...

Tabela 5.4 – Exemplos de categorização de *Word Order* com uma outra etiqueta

ANEXO 5 – EXEMPLOS DE ERROS DE CATEGORIZAÇÃO EM *AGREEMENT* E *WORD ORDER*

Categorias	Nº de dados inseridos nestas categorias	Exemplo				
		Código do segmen-to	Texto de partida	Texto de chegada	Segmentos selecionados pelos anotadores da Unbabel	Propostas de tradução
Word Order / Agreement / Addition	1	W55	*It is possible to purchase our DJ products...*	*É possível comprar [produtos] [do] [nosso] [DJ]...*	produtos / do / nosso / DJ	....comprar nossos produtos DJ...
Word Order / Agreement / Omitted Preposition / Spelling	1	W25	*... workout minimum/ maximum durations...*	*...[treino] [mínimo] / [duração] [máxima]...*	duração /máxima/ mínimo/ treino	...durações mínimas e máximas de treino... OU ...durações mínimas/ máximas de treino...
Word Order / Diacritics / Other POS Omitted	1	W4	*... NINO Letters (confirmed NINO only)...*	*... Nino Letters ([confirmado] [Nino])...*	confirmado/ Nino	...NINO confirmado...
Word Order / Omitted Determiner / Omitted Preposition / Overly Literal	1	W38	*... a booking with a flight provider...*	*... uma reserva [com] passagem aérea [fornecedor]...*	com /fornecedor	...com um fornecedor de passagem aérea...
Word Order / Omitted Preposition / Omitted Determiner / Diacritics	1	W35	*Full page screenshot of the error message...*	*[Página completa] [imagem de captura de tela] do mensagem de erro...*	imagem de captura de tela/Página completa	Imagem de captura de tela da página completa...
Word Order / Omitted Preposition / Overly Literal	1	W48	*... cancel any card payments,...*	*...cancelar qualquer [cartão] [pagamentos],...*	pagamentos/ cartão	...quaisquer pagamentos com cartão...

Tabela 5.5 – Exemplos de categorização de *Word Order* com mais de uma etiqueta

ANEXO 5 – EXEMPLOS DE ERROS DE CATEGORIZAÇÃO EM *AGREEMENT* E *WORD ORDER*

Categorias	**Exemplo**				
	Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Propostas de tradução**
Possivel-mente não natural / Duplication	W42	*...a callback option, please use that...*	*...uma opção de regresso, utilize essa opção, [por favor] [utilize] isso,*	por favor/ utilize	...por favor, utilize isso.../ ...utilize isso, por favor...
Possivel-mente não natural / Punctuation / Diacritics	W63	*...to pay for the order by a (CARD BRAND)/(CARD BRAND) bank card directly.*	*... pagar a encomenda [através] de um cartão bancário (CARD BRAND) / (CARD BRAND). [Diretamente].*	Diretamente/ através	...pagar pela encomenda diretamente através de um cartão bancário... / ...pagar diretamente pela encomenda através de um cartão bancário... / ...pagar pela encomenda através de um cartão bancário diretamente...
Possivel-mente não natural / Word Order / Omitted Preposition	W99	*...contacting the flight provider directly...*	*...entrar em contato [com] [passagem] aérea [fornecedor] [diretamente]...*	diretamente /com/ fornecedor/ passagem	...entrar em contato com o fornecedor de passagem aérea diretamente... / ...entrar em contato diretamente com o fornecedor de

					passagem aérea... / ...entrar diretamente em contato com o fornecedor de passagem aérea...
Possivel-mente não natural / Lexical Selection / Untrans-lated / Ambiguous Translation	W5	... (Drivers Licence must be issued by the same country that you currently reside in)	... (licença de Driais deve ser emitida pelo mesmo país que [atualmente] [está])...	está/ atualmente	...em que você reside atualmente... / ...em que você atualmente reside...

Tabela 5.6 – Exemplos de "Possivelmente não natural" com outras etiquetas

Categoria	**Exemplo**			
	Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**
Overly Literal / Lexical Selection / Ambiguous Translation / Wrong Preposition	W22	*... we'll be able to raise a payment request to our Finance team for you.*	*... poderemos [levantar] um pedido de pagamento [para] a nossa equipe de Finanças para [você].*	levantar/para/ você
Ponctuation / Diacritics	W37	*Hi (PERSON'S NAME F),*	[(PERSON'S NAME F)] [Olá],	Olá/(PERSON'S NAME F)
Lexical Selection / Overly Literal	W101	Wishing you a happy festive season ahead!	Desejando-lhe uma feliz [temporada] [festiva] à frente!	festiva/ temporada

Tabela 5.7 – Exemplos de dados em que há outros tipos de erros (*Word Order*)

Anexo 6
Processos de filtragem: "Separação" e "Adição" de dados

A etapa de "Separação" consistiu essencialmente na divisão de alguns dados que foram previamente selecionados pelos anotadores da Unbabel em duas ou mais linhas, permitindo assim a retificação de segmentos que foram mal categorizados ou a observação mais minuciosa das categorias gramaticais ou funções sintáticas envolvidas no erro. Para exemplificar, na tabela 6.1 a seguir, o dado A148 possui duas expressões assinaladas na segmentação feita pelos anotadores da Unbabel, de categorias gramaticais distintas: o conjunto *preposição* + *artigo definido* [no] e o *possessivo* [seu]. Tendo em vista o interesse da presente pesquisa em observar as categorias gramaticais envolvidas em erros de concordância internamente ao SN, esse dado foi subdividido em duas linhas: o termo [no] foi considerado somente na linha A148 e o termo [seu] foi analisado na segunda linha, A149. Note-se que não se considerou necessário repetir o dado para incluir o núcleo do SN na contagem, pois, nesse caso, ele é o controlador do traço de concordância.

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Segmentos considerados na análise**	**Processos de filtragem (origem do texto)**
A148	*...any funds to remain in your account ...*	*...qualquer fundo continue [no] seu conta...*	no/seu	no	-
A149	*...any funds to remain in your account ...*	*....qualquer fundo continue no [seu] conta...*	-----	seu	Separação (Agreement)

Tabela 6.1 – Etapa de Separação de dados devido às diferentes categorias gramaticais

ANEXO 6 – PROCESSOS DE FILTRAGEM: "SEPARAÇÃO" E "ADIÇÃO" DE DADOS

A tabela 6.2 a seguir apresenta outro exemplo interessante desse processo de Separação, dessa vez devido à segmentação inadequada: em W13, os termos [instituições/financeiras/pré] foram assinalados como pertencentes à categoria de erro *Word Order* pelo anotador da Unbabel. Todavia, as palavras [instituições] e [financeiras] não apresentam nenhum erro, logo não interessam para a presente análise. Já a unidade [pré] foi mal segmentada, pois é na palavra "pré-pago", na sua totalidade, que se verifica um erro de ordem e um erro de concordância. Por isso, durante a análise da presente seção, nenhum segmento foi considerado em W13, tendo em vista a má segmentação feita pelos anotadores. Além disso, através do mencionado processo de Separação, foi criada a linha W14, na qual a unidade "pré" foi separada dos outros segmentos originalmente selecionados pelo anotador da Unbabel e considerada na sua totalidade: [pré-pago].

Código do segmento	**Texto de partida**	**Texto de chegada**	**Segmentos selecionados pelos anotadores da Unbabel**	**Segmentos considerados na análise**	**Processos de filtragem (origem do texto)**
W13	*...(not e-money/pre-paid financial institutions);*	*... (não e-money / pré-pago instituições financeiras);*	instituições/ financeiras/ pré	-	-
W14	*...(not e-money/pre-paid financial institutions);*	*... (não e-money / [pré-pago] instituições financeiras);*	instituições/ financeiras/ pré	pré-pago	Separação (Word Order)

Tabela 6.2 – Etapa de Separação de dados devido à má segmentação

Tendo em vista a busca por uma maior diversidade de estruturas para a análise gramatical da presente seção, preferiu-se examinar para além das unidades selecionadas pelos anotadores da Unbabel. À vista disso, os textos de chegada foram considerados na

íntegra e foram encontrados erros em outras unidades não selecionadas pelos anotadores da Unbabel. Através do processo de "Adição" foi feita a inclusão de dados com expressões que não foram selecionados por esses anotadores, mas que deveriam ter sido, pois envolviam erros de ordem de palavras ou de concordância. Na tabela 46, o segmento [seu] no dado A68 foi originalmente anotado como contendo erro de *Agreement* pelo anotador da Unbabel. Porém, durante a leitura do texto de chegada foi possível encontrar um erro de ordem em outra parte do texto: "*a business account*" deveria ter sido traduzido por "*uma conta empresarial*", mas foi traduzido por "*uma empresa conta*" no texto de chegada. Logo, há erro de ordem, pois o modificador "empresarial" deveria ter sido posicionado após o núcleo do SN, "conta". Em vista disso, esse erro foi adicionado à lista de erros de ordem de palavras através do processo de "Adição", como é possível verificar também no exemplo W49, apresentado igualmente na tabela 6.3.

ANEXO 6 – PROCESSOS DE FILTRAGEM: "SEPARAÇÃO" E "ADIÇÃO" DE DADOS

Código do segmento	Texto de partida	Texto de chegada	Segmentos selecionados pelos anotadores	Segmentos considerados na análise	Processos de filtragem (origem do texto)
A68	*If you signed up for a business account within past 7 days or haven't had much activity on your account...*	*Se você se inscreveu para uma* ***empresa conta*** *nos últimos 7 dias ou não teve muita atividade em [seu] conta...*	seu	seu	-
W49	*If you signed up for a business account within past 7 days or haven't had much activity on your account...*	*Se você se inscreveu para uma [****empresa****] **conta** nos últimos 7 dias ou não teve muita atividade em seu conta...*	-	empresa	Adição (Agreement)

Tabela 6.3 – Etapa de Adição de dados

Anexo 7
Exemplos de sugestões para a Segmentação

São apresentados exemplos da aplicação das sugestões (S1) à (S3) da seção 7.1.1.2.1 na tabela 7.1 a seguir. Os exemplos foram elaborados a partir de excertos já apresentados na *Annotation Guidelines* ou dos dados de tradução fornecidos pela empresa. Nessa tabela, além da demonstração da segmentação adequada segundo as ideias já discutidas anteriormente, são assinalados os traços envolvidos no erro e as categorias gramaticais dos elementos incorretos. No caso dos exemplos EA1* e EA2*, feitos a partir de exemplos apresentados nessas *Guidelines*, os textos de partidas em PB são propostas de tradução feitas no contexto da presente pesquisa. Note-se que, em EA2*, o texto de partida em PB possibilita verificar que o erro está no verbo "*were*" e não no sujeito "*man*", tendo em vista o sujeito ser singular no original.

ANEXO 7 – EXEMPLOS DE SUGESTÕES PARA A SEGMENTAÇÃO

Aplicação das sugestões para segmentação de *Agreement*						
Cód.	**Origem do exemplo**	**Texto de partida**	**Texto de chegada**	**Segmentação adequada**	**Traços envolvidos no erro**	**Categorias gramaticais dos elementos selecionados**
EA1	*Annotation Guidelines* (com adaptações)	Ela informou que a **quantia** cobrada no seu cartão de crédito pelo hotel **foi** um erro.*	*She advised that the **amount** charged on your credit card by the hotel **were** a mistake.*	[were] -> *Agreement*	Número	Verbo
EA2	*Annotation Guidelines* (com adaptações)	O **homem** que eles viram na sexta **era** muito grande.*	*The **man** whom they saw on Friday **were** very big.*	[were] -> *Agreement*	Número	Verbo
EA3	Dado fornecido pela empresa	*...you are trying to create **a new** account...*	…você está tentando criar **um** conta **novo**…	[um] [novo] -> *Agreement*	Gênero	Artigo indefinido e adjetivo
EA4	Dado fornecido pela empresa	*...reply to **this** email with the dates of **the** workouts...*	…responda a **esta** e-mail com as datas **da** treinos…	[esta] -> *Agreement* [da] -> *Agreement*	Gênero e Número	Demonstrativo e artigo definido

ANEXO 7 – EXEMPLOS DE SUGESTÕES PARA A SEGMENTAÇÃO

EA5	Dado forneci-do pela empr-esa (com adapta-ções)	*If you have **credits** you may also use **it** to purchase the element.*	Se você tem **créditos**, você também pode usá-**lo** para comprar o elemento.	[lo] -> *Agree-ment*	Número	Clítico
EA6	Dado forneci-do pela empre-sa (com adapta-ções)	*If you **have** credits you may also use it to purchase the element.*	Se você **tens** créditos, você também pode usá-lo para comprar o elemento.	[tens] -> *Agree-ment*	Pessoa	Verbo

Tabela 7.1 – Exemplos de aplicação das sugestões (S1, S2 e S3)

Também na tabela 7.2 a seguir, foram inseridos exemplos da aplicação das sugestões (S4) à (S6) da seção 7.1.1.2.2 criados a partir da adaptação de excertos já presentes nas *Annotation Guidelines* ou dentre os dados fornecidos pela empresa. Como na tabela anterior, os textos de partida em EW1*, EW2* e EW3* são propostas feitas para o contexto desta subseção. Note que os excertos em EW2 exemplificam um caso em que o primeiro elemento na ordem incorreta (*a small*) é mais extenso em número de palavras do que a palavra vizinha também mal posicionada (*only*), logo deve ser selecionada a segunda unidade. Já no exemplo EW1, ambas as palavras possuem a mesma extensão, logo é escolhida a primeira que ocorreu no texto (*lernen*).

ANEXO 7 – EXEMPLOS DE SUGESTÕES PARA A SEGMENTAÇÃO

Aplicação das sugestões para segmentação de *Word Order*				
Cód.	**Origem do exemplo**	**Texto de partida**	**Texto de chegada**	**Segmen-tação adequada**
EW1	*Annotation Guidelines* (com adaptações)	You have to **learn German** fast.*	*Du musst schnell **lernen Deutsch.***	[lernen] –> Word Order
EW2	*Annotation Guidelines* (com adaptações)	Nós usamos **somente um pequeno** número dos IPs listados nesta página.*	We use **a small only** number of the IPs listed on this page.	[only] -> Word Order
EW3	*Annotation Guidelines* (com adaptações)	Nós também compartilhamos **diariamente** dicas **de design** nas redes sociais.*	We also share **design daily** tips on social media	[design] -> Word Order
EW4	Dado fornecido pela empresa (com adaptações)	Please send them a **billing ticket**.	Por favor envie a eles um **cobrança ingresso**.	[cobrança] -> Word Order
EW5	Dado fornecido pela empresa (com modificações)	We would recommend contacting the flight **provider directly**.	Recomendamos **diretamente** entrar em contato com passagem aérea **fornecedor**.	[direta-mente] -> Word Order [fornece-dor] -> Word Order
EW6	Dado fornecido pela empresa (com modificações)	Tap on the Settings icon at the top right corner.	Toque no ícone Configurações no **direito superior** canto.	[direto] [superior] -> Word Order

ANEXO 7 – EXEMPLOS DE SUGESTÕES PARA A SEGMENTAÇÃO

EW7	Dado fornecido pela empresa (com modificações)	Please send them a billing ticket.	Envie a eles um **por favor** ingresso de cobrança.	[por favor] -> Word Order

Tabela 7.2 – Exemplos de aplicação das sugestões (S4, S5 e S6)

Anexo 8
Exemplos de testes de múltipla escolha para avaliação e treinamento

O presente Anexo 8 propõe exemplos de testes para a avaliação e treinamento dos anotadores e pós-editores da empresa, considerando-se os erros de concordância e ordem de palavras encontrados nos dados fornecidos (cf. seção 6.2) e as dificuldades do processo de anotação dos dados, tendo em vista a complexidade da tipologia de erros utilizada na ferramenta de Anotação da empresa Unbabel (cf. seção 6.1). Os excertos inseridos nas propostas de teste do presente Anexo 8 foram retirados de dados fornecidos pela empresa e já discutidos ao longo da pesquisa a que se refere o presente Anexo (cf. seção 6 e 7).

Cada uma das tabelas apresentadas neste Anexo 8 é dedicada a um assunto distinto, consoante a etapa de anotação trabalhada ou o processo de pós-edição do texto traduzido: testes de segmentação envolvendo concordância (tabela 8.1) e ordem de palavras (tabela 8.2); testes de categorização envolvendo as etiquetas *Agreement* (tabela 8.3), *Word Order* (tabela 8.4) e ambas essas etiquetas (tabela 8.5); testes de conhecimentos linguísticos com uma só respostas correta (tabela 8.6) e com mais de uma resposta correta (tabela 8.7).

O objetivo dos testes nas tabelas (8.1) e (8.2) é verificar se os anotadores dominam as orientações de segmentação das *Annotation Guidelines* e das sugestões (S1) a (S6) apresentadas na seção 7.1 da pesquisa a que se refere o presente Anexo 8, principalmente em casos potencialmente confusos, como a seleção de unidades com o mesmo tipo de erro (T2, T3, T8 e T9), a segmentação de clíticos (T9) e de blocos de palavras (T10), por exemplo.

Procura-se, a partir dos testes de categorização, verificar se os anotadores dominam as diferentes etiquetas utilizadas na ferramenta de anotação da empresa e as orientações mais específicas ligadas a essas etiquetas, principalmente nos contextos que podem causar confusões, como as diferenças entre *Agreement*, *Number* e *Tense/Mood/Aspect* (T14 e T16), bem como a distinção entre *Overly Literal* e *Word Order* (T15). Vale ressaltar que, nesses testes, podem ser selecionadas duas etiquetas para cada unidade, já se adaptando à versão atualizada das *Annotation Guidelines* (cf. seção 3).

O objetivo dos testes de conhecimentos linguísticos nas tabelas (8.6) e (8.7) é verificar o domínio dos anotadores e pós-editores acerca do funcionamento da concordância e da ordem de palavras em PB com excertos e problemas de tradução

próximos aos textos traduzidos pela empresa, como, por exemplo, a concordância entre nomes e modificadores (T19), sujeito e verbo (T21) e no caso de estrangeirismos (T18), bem como a ordem de clíticos (T29 e T30) e advérbios (T28). Como já foi discutido na seção 7.2 da pesquisa, os testes tratando de conhecimentos linguísticos foram divididos em duas categorias: na tabela (8.6) os fenômenos tratados possuem um funcionamento mais constante e, por isso, há somente uma solução possível; já nos exemplos da tabela (8.7) há mais de uma alternativa possível, pois é difícil escolher somente uma forma de traduzir, tendo em vista o funcionamento dos fenômenos abordados nos exemplos.

As perguntas gerais que devem ser respondidas através das alternativas de cada teste foram inseridas logo antes das tabelas a que se referem. Essas tabelas possuem a seguinte organização: na primeira coluna, há o código de identificação do teste; na segunda, são inseridos o excerto no texto de partida, a tradução desse excerto no texto de chegada (no caso dos testes de anotação) e as múltiplas alternativas do teste; na terceira, são apresentadas a sugestão de resolução e uma mensagem de treinamento para os casos em que foram escolhidas as respostas inadequadas. As mensagens de treinamento fornecem mais contexto acerca do tipo de problema contemplado em cada teste.

Pergunta para as Tabelas 8.1 e 8.2 (Testes T1 à T10):

"A partir da observação do texto de partida, escolha a melhor alternativa de segmentação e categorização do(s) erro(s) presente(s) na tradução em PB.

A segmentação será representada através do símbolo '[]' e a categorização através do símbolo '/'. Para exemplificar, em (T0a), as palavras 'um' e 'lindo' foram segmentadas individualmente e categorizadas juntas na mesma etiqueta; em (T0b), as palavras 'um' e 'lindo' foram segmentadas e categorizadas separadamente em etiquetas distintas; em (T0c), foi segmentado e categorizado o bloco de palavras 'um lindo':

T0 – um lindo coroa

a) [um] [lindo]
b) [um] / [lindo]
c) [um lindo]"

ANEXO 8 – EX. DE TESTES DE MÚLTIPLA ESCOLHA PARA AVALIAÇÃO E TREINAMENTO

Testes de segmentação envolvendo a *Concordância*		
Cód.	**Exemplo no texto de partida** **Exemplo de tradução em PB** **Alternativas do teste**	**Mensagem de treinamento**
T1	- *Screenshots of the error message* - Imagens de captura de tela do mensagem de erro. a) [do mensagem] **b) [do]** c) [do] [mensagem] d) [o]	**b) Sugestão de resolução** Incluir somente a unidade incorreta. A unidade mínima de seleção é a palavra inteira. Considerando-se que o determinante "o" está unido à preposição "de", toda a unidade [do] deve ser selecionada. Nesse caso, "mensagem" é o controlador da concordância e não foi mal traduzido, logo não deve ser selecionado.
T2	- *You are trying to create a new account* - Você está tentando criar um conta novo. **a) [um] [novo]** b) [um conta novo] c) [um conta] / [conta novo] d) [um] / [novo]	**a) Sugestão de resolução** Considerando-se que fazem parte da mesma unidade sintática, selecione individualmente as unidades incorretas [um] e [novo] e categorize-os juntos na mesma etiqueta, como em (a). Nesse caso, "conta" é o controlador da concordância e não foi mal traduzido, logo não deve ser selecionado.
T3	- *Reply to this email with the dates of the workouts.* - Responda a esta e-mail com as datas da treinos. a) [esta] [da] b) [esta e-mail com as datas da treinos] **c) [esta] / [da]** d) [esta e-mail] / [da treinos]	**c) Sugestão de resolução** Apesar de apresentarem o mesmo tipo de erro, considerando-se que fazem parte de unidades sintáticas distintas, selecione individualmente as unidades incorretas [esta] e [da] e categorize-os separadamente, como em (c). Nesse caso, "e-mail" e "datas" são controladores das concordâncias e não foram mal traduzidos, logo não devem ser selecionados.

T4	- *If you have credits you may also use it to purchase the element.* - Se você tem créditos, você também pode usá-lo para comprar o elemento. **a) [lo]** b) [usá-lo] c) [créditos] [usá-lo] d) [créditos] [lo]	**a) Sugestão de resolução** No caso de verbos e clíticos separados por hífen, também selecione somente o elemento incorreto. No caso do exemplo, somente o clítico [lo] deve ser selecionado, como em (a). Apesar a referência do clítico [lo] ser controlada pelo elemento "créditos", esse último não deve ser selecionado, pois não contém erro.
T5	- *If you have credits you may also use it to purchase the element.* - Se você tens créditos, você também pode usá-los para comprar o elemento. a) [você tens] b) [você] [tens] **c) [tens]** d) [tens] [pode]	**c) Sugestão de resolução** Apesar de "você" ser o controlador da concordância, selecione somente a unidade que contém o erro. Nesse caso, somente o elemento [tens] deve ser selecionado, como em (c).

Tabela 8.1 – Testes de segmentação envolvendo a *Concordância*

ANEXO 8 – EX. DE TESTES DE MÚLTIPLA ESCOLHA PARA AVALIAÇÃO E TREINAMENTO

Testes de segmentação envolvendo a *Ordem de Palavras*		
Cód.	**Exemplo do texto de partida** **Exemplo de tradução em PB** **Alternativas do teste**	**Mensagem de treinamento**
T6	- *Please send them a billing ticket.* - Por favor envie a eles um cobrança ingresso. a) [cobrança ingresso] **b) [cobrança]** c) [cobrança] [ingresso] d) [ingresso]	**b) Sugestão de resolução** Nos casos de erro de ordem, selecione somente a serem do unidade que, ao ser movida, resolveria o problema. No excerto apresentado, a mudança de posição entre as palavras adjacentes "cobrança" e "ingresso" resolveria o problema de ordem. Tendo em vista o mesmo tamanho (em número de palavras), selecione a primeira que ocorre no texto, ou seja, [cobrança], como em (b).
T7	- *We use only a small number of the IPs listed on this page.* - Nós usamos um pequeno somente número dos IPs listados nesta página. a) [um pequeno] b) [um pequeno somente] c) [usamos] [somente] **d) [somente]**	**d) Sugestão de resolução** Nos casos de erro de ordem, selecione somente a unidade que, ao ser movida, resolveria o problema. No excerto apresentado, a mudança de posição entre o bloco "um pequeno" e a palavra "somente" resolveria o problema de ordem. Contudo, selecione [somente], como em (d), pois é a menor (em número de palavras).
T8	- *Tap on the Settings icon at the top right corner.* - Toque no ícone Configurações no direito superior canto. a) [direito superior canto] b) [direito] / [superior] **c) [direito] [superior]** d) [direito superior]	**c) Sugestão de resolução** Selecione somente as unidades que, ao serem reposicionadas, resolveriam o problema de ordem. Nesse caso, considerando-se a ordem de aparição das palavras no excerto, as unidades [direito] e [superior] devem ser selecionadas individualmente e categorizadas juntas na mesma etiqueta, pois fazem parte da mesma unidade sintática.

ANEXO 8 – EX. DE TESTES DE MÚLTIPLA ESCOLHA PARA AVALIAÇÃO E TREINAMENTO

T9	- *We would recommend contacting the airline ticket provider directly.* - Recomendamos entrar em contato com diretamente passagem aérea fornecedor de. **a) [diretamente] / [passagem aérea]** b) [entrar] [diretamente] /[fornecedor][passagem] c)[diretamente] [fornecedor de] d) [entrar em contato com diretamente] / [passagem aérea fornecedor de]	**a) Sugestão de resolução** Selecione somente as unidades que, ao serem reposicionadas, resolveriam o problema de ordem. No excerto apresentado, há erro de ordem em duas unidades sintáticas distintas: na primeira, a mudança de posição entre "entrar em contato com" e "diretamente" resolveria o problema; na segunda, os segmentos "passagem aérea" e fornecedor de" deveriam trocar de posição. No primeiro caso [diretamente] é menor (em número de palavras). No segundo caso, os dois segmentos citados têm o mesmo tamanho. Logo devem ser selecionadas individualmente [diretamente], pois é a menor, e [passagem aérea], pois é a primeira que aparece do texto de chegada. Ambas as unidades devem ser categorizadas separadamente, como em (a).
T10	- *Please send them a billing ticket.* - Envie a eles um por favor ingresso de cobrança. a) [por favor] [Envie] b) [Envie a eles um por favor] **c) [por favor]** d) [por] [favor]	**c) Sugestão de resolução** Selecione somente a unidade que, ao ser reposicionada, resolveria o problema de ordem. Nesse caso, não selecione as palavras individualmente, prefira selecionar o bloco [por favor] inteiro como em (c). Não há erro de ordem em "Envie a eles um", logo essas palavras não devem ser selecionadas.

Tabela 8.2 – Testes de segmentação envolvendo a *Ordem de Palavras*

ANEXO 8 – EX. DE TESTES DE MÚLTIPLA ESCOLHA PARA AVALIAÇÃO E TREINAMENTO

Pergunta para as tabelas 8.3, 8.4 e 8.5 (Testes T11 à T17)[2]:

"A partir da observação do texto de partida, escolha a melhor alternativa de categorização do segmento assinalado na tradução em PB.

A segmentação será representada através do símbolo '[]' e a união de duas etiquetas através do símbolo '+'. Para exemplificar, em (T0), a palavra 'um' foi segmentada individualmente. Na alternativa (T0a), essa unidade será categorizada sob a etiqueta *Agreement*. Na alternativa, (T0b), essa unidade será categorizada sob ambas as etiquetas *Agreement* e *Word Order*.

T0 – lindo coroa [um]

a) Agreement
b) Agreement + Word Order"

Testes de categorização envolvendo a etiqueta *Agreement*		
Cód.	**Exemplo do texto de partida, Exemplo de tradução em PB e Alternativas do teste**	**Mensagem de treinamento**
T11	**-** ***An*** *account from our company is for personal use only.* - [**As**] conta da nossa empresa é somente para uso pessoal. a) Overly Literal b) Agreement **c) Agreement + Wrong Determiner** d) Agreement + Omitted Determiner	**c) Sugestão de resolução** Além do erro de *Agreement* entre o determinante e o núcleo do SN, há o uso incorreto de um artigo definido no lugar de um indefinido, como está no texto de partida. Por isso, ambas as categorias *Agreement* e *Wrong Determiner* devem ser selecionadas, como em (c). Não há omissão do determinante e a tradução não foi feita literalmente, por isso as categorias *Overly Literal* e *Omitted Preposition* não são adequadas para este segmento.
T12	- *Take a look at our Help Center article.*	**a) Sugestão de resolução**

[2] A segmentação foi delimitada pelo símbolo "[]" nos excertos apresentados nos testes.

	- Olhe o nosso artigo do Centro [**da**] Ajuda. **a) Addition** b) Agreement c) Wrong Determiner d) Wrong Preposition	A tradução correta desse excerto é "Olhe o nosso artigo do Centro de Ajuda". Não há uma discordância de traços entre o determinante "a" e o nome "Ajuda", logo a etiqueta *Agreement* não se aplica. Apesar de o erro envolver o elemento correspondente à união "preposição + artigo definido", as etiquetas *Wrong Determiner* e *Wrong Preposition* não podem ser selecionadas, pois o erro se encontra na adição desse artigo definido, tendo em vista ele ser desnecessário nesse contexto.
T13	- *All documents **must** be older than 3 months.* - Todos os documentos [**deve**] ser superiores a 3 meses. **a) Agreement** b) Wrong Auxiliary Verb c) Tense/Mood/Aspect d) Inconsistency	**a) Sugestão de resolução** No exemplo, há discordância entre os traços de número do sujeito e do verbo auxiliar, aplicando-se então a etiqueta *Agreement*, como em (a). Não confundir essa etiqueta com *Wrong Auxiliary Verb*, *Inconsistency* ou *Tense/Mood/Aspect* : o primeiro se refere aos casos em que o verbo auxiliar utilizado é inadequado para o contexto, sendo preferível outra forma; o segundo aos casos em que a texto apresenta inconsistências na tradução do mesmo termo ou no uso de abreviações; o terceiro se refere aos casos em que o verbo apresenta erro no tempo, modo e aspecto.

Tabela 8.3 – Testes de categorização envolvendo a etiqueta *Agreement*

Testes de categorização envolvendo a etiqueta *Word Order*		
Cód.	**Exemplo do texto de partida, Exemplo de tradução em PB e Alternativas do teste**	**Mensagem de treinamento**
T14	- *Do you have a subscription plan?*	**b) Sugestão de resolução**

	- Você possui um [**assinatura**] plano? a) Word Order + Wrong Preposition **b) Word Order + Omitted Preposition** c) Word Order d) Overly Literal	A tradução correta da unidade sintática na qual se encontra a unidade selecionada é "um plano de assinatura". Apesar da influência da ordem do texto de partida no texto de chegada, levando a uma tradução muito literal, prefira categorizar através de etiquetas mais específicas. Há erro de ordem, pois "assinatura" deve estar após o seu núcleo "plano". Não confunda as etiquetas *Wrong Preposition* e *Omitted Preposition*: o primeiro se refere ao uso incorreto de uma preposição; o segundo à falta dessa preposição no texto de partida.
T15	- *An attached screenshot of the issue can help us a lot navigating the problem as well.* - Um anexo da captura de tela do problema pode nos ajudar a entender **muito** o problema também. a) Ambiguous Translation b) Over translated **c) Word Order** d) Addition	**c) Sugestão de resolução** Apesar de o texto de chegada não ser agramatical, há erro de *Word Order* na unidade selecionada, pois sua posição causou uma mudança no sentido do texto original. A melhor posição para o advérbio assinalado é logo após o verbo "ajudar". A etiqueta *Over translated* não se aplica ao erro assinalado, pois o texto de chegada não é mais específico do que o texto de partida. Também a etiqueta *Addition não* é adequada, pois não houve adição de elementos na tradução, mas sim um reposicionamento da unidade selecionada. Finalmente, não há ambiguidade no texto de chegada, logo não se aplica a etiqueta *Ambiguous Translation.*

Tabela 8.4 – Testes de categorização envolvendo a etiqueta *Word Order*

ANEXO 8 – EX. DE TESTES DE MÚLTIPLA ESCOLHA PARA AVALIAÇÃO E TREINAMENTO

Testes de categorização envolvendo as etiquetas *Agreement* e *Word Order*		
Cód.	**Exemplo do texto de partida, Exemplo de tradução em PB e Alternativas do teste**	**Mensagem de treinamento**
T16	- *We don't accept **pre-paid** financial institutions.* - Nós não aceitamos [**pré-paga**] instituições financeiras. a) Word Order + Number **b) Agreement + Word Order** c) Overly Literal d) Agreement	**b) Sugestão de resolução** Há efetivamente erro de concordância entre os traços de número da palavra selecionada ("pré-paga") e do núcleo da unidade sintática em que ela se encontra ("instituições"). Também há erro de ordem, pois a unidade assinalada deveria estar posicionada após "instituições financeiras" em PB. Não confunda as etiquetas *Agreement* e *Number*: o primeiro se refere à concordância de traços, o segundo à inconsistência entre os numerais no texto de partida e chegada. Apesar da influência da ordem do texto de partida no texto de chegada, levando a uma tradução muito literal, prefira categorizar através de etiquetas mais específicas.
T17	- *You signed up for a **business** account.* - Você se inscreveu para uma [**empresa**] conta. a) Word Order + Agreement b) Agreement c) Word Order **d) Word Order + POS**	**d) Sugestão de resolução** A tradução correta desse excerto é "Você se inscreveu para uma conta empresarial". Logo, a palavra "*business*" deveria ser posicionada após o núcleo do SN em PB e deveria ter sido traduzida por um adjetivo correspondente. Por isso, ambas as categorias *Word Order* e *POS* devem ser selecionadas, como em (d).

Tabela 8.5 – Testes de categorização envolvendo as etiquetas *Agreement* e *Word Order*

ANEXO 8 – EX. DE TESTES DE MÚLTIPLA ESCOLHA PARA AVALIAÇÃO E TREINAMENTO

Pergunta para a tabela 8.6 (Testes T18 à T24):

"A partir da observação do texto de partida, escolha uma só alternativa com a opção mais adequada de tradução em PB."

Testes de conhecimentos linguísticos com uma só resposta correta		
Cód.	**Exemplo no texto de partida e Alternativas do teste**	**Mensagem de treinamento de acordo com a resposta escolhida**
T18	*your company is not a Venture Builder* a) sua empresa não é Venture Builder b) sua empresa não é um *Venture Builder* **c) sua empresa não é uma *Venture Builder*** d) sua empresa não é *Venture Builder*	**c) Sugestão de resolução** Faça também a concordância de traços no caso de estrangeirismos. Nesse tipo de palavra é aconselhado assinalar a palavra com o itálico no texto de chegada. Evite omitir os determinantes.
T19	*We don't accept prepaid financial institutions.* a) Não aceitamos pré-paga instituições financeiras. **b) Não aceitamos instituições financeiras pré-pagas.** c) Não aceitamos instituições financeiras pré-paga. d) Não aceitamos pré-pagas instituições financeiras.	**b) Sugestão de resolução** Em PB, os modificadores geralmente se posicionam após o núcleo do SN e devem concordar com ele em número e gênero.
T20	*On your next download, purchase/s will be deducted from your credits.* a) No seu próximo download, a compra será deduzida dos seus créditos.	**d) Sugestão de resolução** No texto original, há uma alternância entre os traços de número, representado por "/". No caso da alternância de

	b) No seu próximo download, as compras serão deduzidas dos seus créditos. c) No seu próximo download, a/ compra/s será/ão deduzida/s dos seus créditos. **d) No seu próximo download, a(s) compra(s) será(ão) deduzida(s) dos seus créditos.**	traços em PB, utilize o símbolo "()".
T21	*All documents must show the issue date and should not be older than 3 months.* a) Todos os documentos devem mostrar a data de emissão e não deve ser superior a 3 meses. **b) Todos os documentos devem mostrar a data de emissão e não devem ser superiores a 3 meses.** c) Todos os documentos deve mostrar a data de emissão e não deve ser superior a 3 meses. d) Todos os documentos devem mostrar a data de emissão e não devem ser superior a 3 meses.	**b) Sugestão de resolução** É importante verificar os traços do referente dos sujeitos ocultos durante a pós-edição. No excerto apresentado, o verbo "devem" e o adjetivo "superior" devem concordar com os traços de [Todos os documentos].
T22	*We know you were immediately charged $10.* a) Sabemos que você foi cobrado imediatamente $10. b) Sabemos que $10 foi imediatamente cobrado de você. **c) Sabemos que $10 foram imediatamente cobrados de você.** d) Sabemos que de você foi imediatamente cobrado $10.	**c) Sugestão de resolução** Evite manter a ordem do texto traduzido demasiadamente próxima à ordem do texto original. Nos exemplos, a tradução literal da voz passiva causa problemas, pois esse tipo de construção é diferente em PB e em inglês.
T23	*Check the workout minimum/maximum durations*.	**c) Sugestão de resolução**

	a) Verifique o treino mínimo/durações máximas. b) Verifique as durações mínima/máxima de treino. c) **Verifique as durações mínimas/máximas de treino.** d) Verifique os treinos mínimos/durações máximas.	É importante reconhecer o núcleo da unidade sintática no texto original para melhor avaliar a existência de erro no texto traduzido. No texto de partida, "durations" é o núcleo do SN, "workout" é o complemento e há uma alternância entre os dois modificadores "minimum/maximum". Na tradução para PB, a ordem desses elementos muda drasticamente, tendo em vista as diferenças entre inglês e PB.
T24	*We make sure that all secret documents are safe.* a) Certificamos que todo o secreto documento está a salvo. b) Certificamos que todos os secretos documentos estão a salvo. c) Certificamos que todo o documento secreto está a salvo. **d) Certificamos que todos os documentos secretos estão a salvo.**	**d) Sugestão de resolução** Os traços do núcleo do SN no original "*documents*" devem ser mantidos no texto de chegada. Além disso, é possível encontrar adjetivos antes do núcleo do SN em PB, mas há certas diferenças semânticas entre os adjetivos que são permitidos e vetados nessa posição. No caso dos excertos apresentados, "secreto" deve ser posicionado após o seu núcleo em PB.

Tabela 8.6 – Testes de conhecimentos linguísticos com uma resposta correta

ANEXO 8 – EX. DE TESTES DE MÚLTIPLA ESCOLHA PARA AVALIAÇÃO E TREINAMENTO

Pergunta para a tabela 8.7 (Testes T25 à T31):

"A partir da observação do texto de partida, escolha as alternativas com as traduções mais adequadas em PB. Lembre-se que nestes casos é possível selecionar mais de uma opção."

Testes de conhecimentos linguísticos com múltiplas respostas corretas		
Cód.	**Exemplo no texto de partida** **Exemplo de tradução em PB** **Alternativas do teste**	**Mensagem de treinamento de acordo com a resposta escolhida**
T25	*You have a month to respond to bank inquiries.* a) Você tem um mês para responder ao banco dos inquéritos. **b) Você tem um mês para responder aos inquéritos bancários.** **c) Você tem um mês para responder aos inquéritos do banco.** d) Você tem um mês para responder aos bancários inquéritos.	**b) e c) Sugestões de resolução** Em PB, nomes não podem modificar ou complementar diretamente o núcleo do SN. Por isso, além da mudança na ordem de "*bank*", é necessário inserir uma preposição (c) ou mudar a categoria gramatical desse nome (a).
T26	*Most visual issues can be resolved by reconnecting your cables.* **a) A maioria dos problemas visuais podem ser resolvidos reconectando os seus cabos.** b) A maioria dos problemas visuais pode ser resolvido reconectando os seus cabos. **c) A maioria dos problemas visuais pode ser resolvida reconectando os seus cabos.** d) A maioria dos problemas visuais podem ser resolvido reconectando os seus cabos.	**a) e c) Sugestões de resolução** Há em PB o fenômeno de concordância semântica. Os excertos "*podem ser resolvidos*/*pode ser resolvidos*" podem concordar gramatical-mente com os traços do núcleo do SN sujeito "maioria" ou concordar semanticamente com o núcleo do SN mais encaixado "problemas".
T27	*This is a tool to help travellers plan their trips and find the best airline.*	**c) e d) Sugestões de resolução** Em certas construções do PB, como a do excerto apresentado, é possível a

	a) Esta uma ferramenta para ajudar viajantes a programarem as suas viagens e encontrar a melhor companhia aérea. b) Esta uma ferramenta para ajudar viajantes a programar as suas viagens e encontrarem a melhor companhia aérea. **c) Esta uma ferramenta para ajudar viajantes a programarem as suas viagens e encontrarem a melhor companhia aérea.** **d) Esta uma ferramenta para ajudar viajantes a programar as suas viagens e encontrar a melhor companhia aérea.**	escolha entre o infinitivo não-flexionado ou o infinitivo flexionado.
T28	*We need to be sure as you indicated in the questionnaire the following:* **a) Precisamos ter certeza, pois você indicou no questionário, o seguinte:** **b) Precisamos ter certeza, pois, o seguinte, você indicou no questionário:** **c) Precisamos ter certeza, pois você indicou o seguinte no questionário:** **d) Precisamos ter certeza, pois você, o seguinte, indicou no questionário:**	**Todas podem ser sugestões de resolução, mas (c) é a mais natural.** Em PB, os constituintes frásicos podem ocupar diversas posições na frase. A opção mais natural foi apresentada em (c), pois segue a ordem básica SVO do PB. Mas as outras opções também são possíveis. No momento da pós-edição, verifique a melhor opção consoante o contexto. Não se prenda à ordem apresentada no texto de partida, procure a ordem mais natural para os falantes do PB.
T29	*We sent an e-mail to help you choose a new password.* a) Nós enviamos um e-mail para ajudar-lhe a escolher uma senha nova.	**b) e c) podem ser sugestões de resolução, mas (c) é a mais natural.** O clítico "lhe" normalmente exerce função

	b) Nós enviamos um e-mail para o ajudar a escolher uma senha nova. c) Nós enviamos um e-mail para lhe ajudar a escolher uma senha nova. **d) Nós enviamos um e-mail para ajudá-lo a escolher uma senha nova.**	de complemento indireto, logo não é adequado neste contexto. Ambas as posições (b) e (c) são possíveis em PB, mas (d) é opção mais natural.
T30	*If you'd like to change to the monthly subscription, please let me know.* **a) Se você gostaria de mudar para a subscrição mensal, me informe.** b) Se você gostaria de mudar para a subscrição mensal, informe a mim. c) Se você gostaria de mudar para a subscrição mensal, informe mim. **d) Se você gostaria de mudar para a subscrição mensal, informe-me.**	**a) e d) podem ser sugestões de resolução, mas (a) é a mais natural.** Tendo em vista a tendência proclítica do PB, a opção (a) é a mais natural. Contudo, ainda é possível encontrar em PB construções como (a).
T31	*An attached screenshot of the issue can help us a lot navigating the problem.* **a) Uma captura de tela anexa do problema pode nos ajudar muito a entender o problema.** b) Uma captura de tela anexa do problema pode nos ajudar a entender muito o problema. **c) Uma captura de tela anexa do problema muito pode nos ajudar a entender o problema.** d) Uma captura de tela anexa do problema pode nos muito ajudar a entender o problema.	**a) e c) podem ser sugestões de resolução, mas (a) é a mais natural.** Os advérbios podem ocupar várias posições em PB. Porém, há posições agramaticais, como entre o clítico e o verbo em (d) e posições gramaticais, mas inadequadas, pois alteram o sentido do texto original como em (b).

Tabela 8.7 – Testes de conhecimentos linguísticos com múltiplas respostas corretas

Anexo 9
Exemplos de *Golden Text* para avaliação e treinamento

No presente Anexo 9 são sugeridos exemplos de textos que podem ser utilizados na elaboração de *Golden Texts* para avaliar o domínio dos anotadores acerca das orientações dadas nas *Annotation Guidelines* (Tabela 9.1), bem como os conhecimentos linguísticos de anotadores e pós-editores da empresa quanto ao funcionamento da ordem de palavras e da concordância em PB (Tabelas 9.1 e 9.3), baseando-se nas dificuldades e erros encontrados durante a análise dos dados feita na seção 6.

As traduções apresentadas nos *Golden Texts* a seguir possuem alguns erros previamente controlados, ligados à uma resposta ideal para a anotação e pós-edição de cada um desses erros. A avaliação pode ser feita através da comparação entre o que foi feito pelas pessoas avaliadas e aquilo que idealmente deveria ter sido feito, presente no *Golden Text*. O treinamento pode ser feito a partir de mensagens enviadas às pessoas avaliadas, caso tenham feito anotações ou edições inadequadas.

Os *Golden Texts* da Tabela 9.1 tratam de ambas as etapas de segmentação e categorização de erros. Durante a avaliação, podem ser selecionadas duas etiquetas para cada unidade, conforme a nova versão das *Annotation Guidelines* (cf. seção 3), e foram consideradas as sugestões (S1) à (S7) apresentadas na seção 7.2 da pesquisa a que se refere o presente anexo. As colunas dessa Tabela 9.1 apresentam a seguinte organização: na primeira coluna, há o código do *Golden Text*; na segunda, há o texto original na língua de partida; na terceira, está inserida a tradução na língua de chegada, contendo os erros controlados; na quarta, são fornecidas as formas adequadas de segmentar e categorizar os erros presentes no texto de chegada. Não foram inseridas propostas de *feedback* na Tabela 9.1 tendo em vista os excertos e erros dessa tabela serem muito próximos aos testes de múltipla escolha das Tabelas 8.1 a 8.5 do Anexo 8 e, consequentemente, poderem ser utilizadas as mesmas mensagens de treinamento.

No caso dos *Golden Texts* das Tabelas 9.2 e 9.3, os erros de tradução inseridos envolvem alguns dos fenômenos de concordância e ordem de palavras em PB tratados nas seções 5, 6 e 7 da pesquisa a que se refere o presente Anexo 9. Os fenômenos apresentados na Tabela 9.2 são mais regulares, em que é possível identificar regras fixas acerca do funcionamento da concordância e da ordem de palavras, por isso foi inserida somente uma proposta de tradução em cada *Golden Text*. Já no caso da Tabela 9.3, os fenômenos são mais problemáticos, sendo difícil determinar uma única forma de tradução, por isso foram inseridas duas propostas de tradução para cada *Golden Text,*

sendo ainda possível criar mais possibilidades de tradução envolvendo os fenômenos tratados. Nessas tabelas, são inseridas também mensagens de *feedback* para auxiliar na etapa de treinamento das pessoas avaliadas. Essas duas tabelas se organizam de maneira similar à Tabela 9.1: há o código do *Golden Text*, o texto original na LP, a tradução na LC com os erros controlados, a(s) proposta(s) de *Golden Text* a ser(em) comparado(s) com edição feita pela pessoa avaliada e uma mensagem de treinamento, caso a pós-edição tenha sido feita de forma inadequada.

Orientação para a Tabela 9.1 (*Golden Texts* GT1 à GT5)[3]:

"A partir da observação do texto de partida, anote os erros presentes na tradução em PB."

Cód.	Texto de partida	Tradução na língua de chegada	Anotação adequada dos segmentos
GT1	If you have a problem with the flight, we would recommend contacting the airline ticket provider directly. You can also take a look at our Help Center article.	Se você tens problemas com o voos, recomendamos entrar em contato com diretamente passagem fornecedor aérea.	[tens] -> Agreement [voos] -> Spelling [diretamente] -> Word Order [passagem] -> Word Order [fornecedor] -> Omitted Preposition
GT2	You are trying to create a new *Exclusive* account but before you signed up for a business account. We would like to remind you that an *Exclusive* account from	Você está tentando criar um conta novo *Exclusive*, mas antes você se inscreveu para uma empresa conta. Gostaríamos de lembrá-lo que as conta *Exclusive* da nossa empresa é	[um] [novo] -> Agreement [novo] -> Agreement + Word Order [empresa] -> Word Order + POS [as] -> Word Order + Wrong Determiner

[3] A anotação apresentada na última coluna foi representada através de símbolos: a segmentação das unidades foi delimitada por "[]"; a categorização das unidades foi representada por "->"; e a união de duas etiquetas para a mesma unidade foi representada por "+".

	our company is for personal use only.	somente para uso pessoal.	
GT3	We don't accept pre-paid financial institutions. If you have credits you may also use it to purchase the element. You can do this by tapping on the *Settings* icon at the top right corner.	Nós não aceitamos pré-paga instituições financeiras. Se você tens créditos, você também pode usá-las para comprar o elemento. Você podes fazer isso tocando no ícone *Configurações* no direito superior canto.	[pré-paga]-> Word Order [tens] -> Agreement [las] -> Agreement [podes] -> Agreement [direito] [superior]-> Word Order
GT4	We use only a small number of the IPs listed on this page. To know more, take a look at our Help Center article.	Nós usamos um pequeno somente número dos IPs listados nesta página. Para saber mais, olhe o nosso artigo do Centro da Ajuda.	[somente] -> Word Order [IP] -> Spelling [a nossa] -> Agreement [da] -> Addition
GT5	Reply to this email with the dates in the workouts. An attached screenshot of the issue can help us a lot navigating the problem as well.	Responda a esta e-mail com as datas da treinos. Um captura de tela anexo do problema pode nos ajudar a entender muito o problema também.	[esta] -> Agreement [da] -> Agreement + Wrong Preposition [anexo] -> Word Order + Omitted Preposition [muito] -> Word Order

Tabela 9.1 – *Golden Text* para a avaliação e treinamento de anotadores

ANEXO 9 – EXEMPLOS DE *GOLDEN TEXT* PARA AVALIAÇÃO E TREINAMENTO

Orientação para as Tabelas 9.2 e 9.3 (*Golden Texts* GT5 à GT8):

"A partir da observação do texto de partida, edite os erros presentes na tradução em PB."

***Golden Text* para a avaliação e treinamento de editores e anotadores**				
Cód.	**Texto de partida**	**Texto de chegada**	***Golden Text***	**Mensagem de treinamento**
GT5	*We know you were immediately charged $10. As we don't accept prepaid financial institutions, on your next download, purchase/s will be deducted from your credits.*	Sabemos que você foi cobrado imediatamente $10. Como não aceitamos pré-paga instituições financeiras, no seu próximo download, a compra/s será deduzida dos seus créditos.	Sabemos que **$10 foram cobrados** imediatamente **de você**. Como não aceitamos instituições financeiras **pré-pagas**, no seu próximo download, **a(s) compra(s) será(ão) deduzida(s)** dos seus créditos.	Evite manter a ordem do texto traduzido demasiadamente próxima à ordem do texto original. Nos exemplos, a tradução literal da voz passiva causa problemas, pois esse tipo de construção é diferente em PB e em inglês. No caso da alternância de traços em PB, utilize o símbolo "()".Em PB, os modificadores geralmente se posicionam após o núcleo do SN e devem concordar com ele em número e gênero.
GT6	*Remember that all documents must show the issue date and should not be older than 3 months. And don't worry:*	Lembre-se que todos os documentos deve mostrar a data de emissão e não deve ser superior a 3 meses. E não se preocupe:	Lembre-se que todos os documentos **devem** mostrar a data de emissão e não **devem** ser **superiores** a 3 meses. E não se preocupe:	É importante verificar os traços do referente dos sujeitos ocultos durante a pós-edição. No excerto apresentado, o verbo "devem" e o adjetivo "superior" devem concordar com os traços de [Todos os documentos]. Os traços do núcleo do SN

	we make sure that all secret documents are safe.	certificamos que todo secreto documento está a salvo.	certificamos que **todos os documentos secretos estão** a salvo.	no original "*documents*" devem ser mantidos no texto de chegada. Além disso, é possível encontrar adjetivos antes do núcleo do SN em PB, mas há certas diferenças semânticas entre os adjetivos que são permitidos e vetados nessa posição. No caso dos excertos apresentados, "secreto" deve ser posicionado após o seu núcleo em PB.

Tabela 9.2 – *Golden Text* para a avaliação e treinamento de editores e anotadores (1)

ANEXO 9 – EXEMPLOS DE *GOLDEN TEXT* PARA AVALIAÇÃO E TREINAMENTO

***Golden Text* para a avaliação e treinamento de editores e anotadores**					
Cód.	**Texto de partida**	**Texto de chegada**	***Golden Text* – possível tradução 1**	***Golden Text* – possível tradução 2**	**Mensagem de treinamento**
GT7	*An attached screenshot of the issue can help us a lot navigating the problem, but most visual issues can be resolved by reconnecting your cables.*	Uma captura de tela anexa do problema pode nos ajudar a entender muito o problema, mas a maioria dos problemas visuais pode ser resolvido reconectando os seus cabos.	Uma captura de tela anexa do problema pode nos ajudar **muito** a entender o problema, mas a maioria dos problemas visuais **podem ser resolvidos** reconectando os seus cabos.	Uma captura de tela anexa do problema **muito** pode nos ajudar a entender o problema, mas a maioria dos problemas visuais **pode ser resolvida** reconectando os seus cabos.	Os advérbios podem ocupar várias posições em PB. Porém, na ordem apresentada o advérbio "muito" altera o sentido do texto original. Há em PB o fenômeno de concordância semântica. Logo, é possível a concordância gramatical com os traços do núcleo do SN sujeito "maioria" ou semântica com o núcleo do SN mais encaixado "problemas".
GT8	*We sent an e-mail to help you choose a new*	Nós enviamos um e-mail para ajudar a lo escolher uma senha	Nós enviamos um e-mail para **o ajudar** a escolher uma senha	Nós enviamos um e-mail para **ajudá-lo** a escolher uma senha	Ambas as posições "o ajudar" e "ajudá-lo" são possíveis em PB,

	password. If you'd like to change to the monthly subscription, please let me know	nova. Se você gostaria de mudar me para a subscrição mensal, informe.	nova. Se você gostaria de mudar para a subscrição mensal, **me informe**.	nova. Se você gostaria de mudar para a subscrição mensal, **informe-me**.	mas a última é opção mais natural. No caso do clítico "me", ambas as expressões "informe-me" e "me informe" são possíveis, mas a última é a mais natural.

Tabela 9.3 – *Golden Text* para a avaliação e treinamento de editores e anotadores (2)